

国际信息工程先进技术译丛



全IP网络融合

Convergence Through All-IP Networks

[美] 阿索克 K. 塔卢克达尔（Asoke K. Talukder）
努诺 M. 加西亚（Nuno M. Garcia）
加雅希尔莎 G. M.（Jayateertha G. M.） 主编

王玲芳 刘洋 等译









机械工业出版社
CHINA MACHINE PRESS

国际信息工程先进技术译丛

全IP网络融合

Convergence Through All-IP Networks

阿索克 K. 塔卢克达尔（Asoke K. Talukder）
［美］ 努诺 M. 加西亚（Nuno M. Garcia） 主编
加雅希尔莎 G. M. （Jayateertha G. M.）
王玲芳 刘 洋 等译



机 械 工 业 出 版 社

图书在版编目（CIP）数据

全 IP 网络融合/（美）塔卢克达尔（Talukder，A. K.）等主编；王玲芳等译. —北京：机械工业出版社，2016.2
（国际信息工程先进技术译丛）
书名原文：Convergence Through All-IP Networks
ISBN 978-7-111-52645-2

Ⅰ.①全…　Ⅱ.①塔…②王…　Ⅲ.①计算机网络－通信协议　Ⅳ.①TN915.04

中国版本图书馆 CIP 数据核字（2016）第 001597 号

机械工业出版社（北京市百万庄大街 22 号　邮政编码 100037）
策划编辑：张俊红　责任编辑：闾洪庆
责任校对：陈　越　封面设计：马精明
责任印制：李　洋
三河市国英印务有限公司印刷
2016 年 3 月第 1 版第 1 次印刷
169mm×239mm · 25 印张 · 516 千字
标准书号：ISBN 978-7-111-52645-2
定价：99.00 元

凡购本书，如有缺页、倒页、脱页，由本社发行部调换

电话服务	网络服务
服务咨询热线：010-88361066	机工官网：www.cmpbook.com
读者购书热线：010-68326294	机工官博：weibo.com/cmp1952
010-88379203	金书网：www.golden-book.com
封面无防伪标均为盗版	教育服务网：www.cmpedu.com

本书共 11 章，系统讲解了下一代网络（NGN）技术全部体系，内容涵盖骨干网络、各种网元以及大量的用户端设备。书中解释了全 IP 网络及 IP 融合，包含带有专用服务和应用场景的各种专题，具体内容涉及从光纤骨干网到无线“最后一英里”、物联网、低功率无线个域网（LoWPAN）和扩展的家庭联网、WiMAX、常用路由及 IPv6 路由、高速路上的车联网以及车内/车间的通信、NGN 中的网络安全问题等多个专题。

本书适合互联网领域和信息通信、移动通信等领域的专业研发和设计人员阅读，也可作为业界从业人员以及高等院校计算机和信息通信等专业高年级师生的参考书。



译 者 序

网络融合的概念，国内从20世纪90年代开始就有人在提，原来的说法是“三网合一”。经过近20年的技术发展，原来设想的以一种网络技术替代所有其他技术的想法，现在看来似乎有点理想化了，所以在现实中也就没有出现。在21世纪初，人们又提出了“网络融合”的说法，其大意是指将不同的网络技术在提供应用和服务方面融合，更确切地说就是体验融合、应用融合。本书则从另一个角度，即网络层的角度提出了“全IP网络融合”的理念。事实证明，因特网的理念和实施技术，也确实在网络层将多样化的物理层技术融合在了一起。如今网络界所开展的未来网络的研究，其目标多是单纯的、对某种网络技术的研究，而没有考虑到网络互联的融合问题。本书提出的“全IP网络融合”理念，是否能够适用于未来场景，让我们大家共同拭目以待吧。

本书共11章，系统讲解了下一代网络（NGN）技术全部体系，内容涵盖骨干网络、各种网元以及大量的用户端设备。书中解释了全IP网络及IP融合，包含带有专用服务和应用场景的各种专题，具体内容涉及从光纤骨干网到无线“最后一英里”、物联网、低功率无线个域网（LoWPAN）和扩展的家庭联网、WiMAX、常用路由及IPv6路由、高速路上的车联网以及车内/车间的通信、NGN中的网络安全问题等多个专题。

本书由王玲芳负责第1章、第2章、第9~11章的翻译以及全书的统稿和校对工作，刘洋负责第3~8章的翻译工作。本书在翻译过程中，李虹、潘东升、李冬梅、吴秋义、王弟英、吴璟、游庆珍、李传经、王领弟、王建平、李睿、吴昊、王灵芹、张永、李志刚、左会高、申永林、潘贤才、刘敏、李钰琳、王青改、李倩、陈军、许侠林、王改玲、张增军、李岩、冯佰永等同志也参加了部分翻译工作，在此表示感谢。同时感谢机械工业出版社的编辑和相关同志。

不过，需要特别指出的是，本书的内容仅代表作者个人的观点和见解，并不代表译者及其所在单位的观点。另外，由于翻译时间比较仓促，译文中疏漏错误之处在所难免，敬请广大读者原谅并批评指正。

译 者

原书前言

一项技术的成功是以对一名用户来说有多么不可见（即用户感知不到技术的存在）来度量的。21世纪是任何地方可通信的世纪——任何人可在任何时间、世界上的任何地方如此容易、如此无缝地进行通信，不管通信传输的是话音、数据、多媒体还是视频，均是如此。虽然对一名用户而言，这看起来是简单的，但从科学和工程角度来看，存在复杂的网络结构和技术，它们在后台协同工作，共同产生这些增效和奇迹。要解释这些奇迹的互联，是写作本书背后的动机。

2009年4月28日到30日，我们在开罗举办了第6届IEEE和IFIP无线和光通信网络国际会议（WOCN2009）。在该会议中，Talukder博士就下一代网络（NGN）做了一次辅导讲座。之后开始了本书的奠基性工作——Talukder博士和Garcia博士在那次会议相遇，Pan Stanford出版社的Stanford Chong先生与Talukder博士探讨就所说专题撰写一部书的事情，后来Jayateertha博士加入了这个团队。

我们的目标是写出一本书，涵盖NGN全技术谱系，从骨干到各种网元，以及大量的用户端设备。我们想要这样一本书，它解释了全IP及其融合，融合以另外的方式保持不可见。就这个方面而言，本书包括各种专题，其中有专用服务和应用场景。在这样做时，我们的主要活动是尽量将这些复杂专题介绍给读者，在呈现方面不失简单性和可读性。我们想要针对工业界的一本完备的手册及针对学生、专业人员和研究人员的一本参考书。

为取得融合和NGN的上述目标，本书包括从光纤骨干到无线最后一英里等各种专题，其中还有路由；包括“物联网”、低功率无线个域网（LoWPAN）和扩展的联网家庭；包括移动性和全球微波接入互操作性（WiMAX）；包括路由，广泛涉及了IPv6路由；包括高速路上的车用网络以及车内和车间通信。在21世纪，有关网络的一本书，如果没有涉及安全问题，则是不完全的。因此，本书也包括NGN中的安全问题。

使具有这样一个宽范围谱系专题（涵盖下一代互联网和融合）的一本书有其自身的挑战。挑战的最困难部分是找到可撰稿的专家和作者的合适组合。虽然这花费了大量时间，但幸运的是，我们得到世界级的学术带头人参与进来作为本书的作者。我们尽力使本书没有错误，并尊重原作者以及商标和版权。但是，任何不经意的错误或疏忽都是令人觉得遗憾的。

我们真诚地感谢所有撰稿人，特别感谢Pan Stanford出版社，感谢它即将出版

本书。我们感谢审稿人和编辑团队的辛勤努力，奉献出一部卓越的书。我们也感谢所有作者和编辑的所有家庭成员，感谢他们的支持。

Asoke K. Talukder
Nuno M. Garcia
Jayateertha G. M.

目　录

译者序

原书前言

第1章　全IP网络：导论 …… 1

1.1　各代因特网 …… 1

1.2　无线因特网 …… 2

1.3　全IP网络 …… 6

参考文献 …… 8

第2章　IPv6中的寻址和路由 …… 9

2.1　引言 …… 9

2.2　寻址 …… 9

2.2.1　寻址概述 …… 10

2.2.2　单播寻址 …… 14

2.2.3　组播寻址 …… 20

2.2.4　任意播地址 …… 23

2.2.5　主机和路由器的地址 …… 24

2.2.6　地址块分配 …… 25

2.2.7　单播或任意播地址指派规程 …… 27

2.3　IPv4到IPv6转换 …… 31

2.3.1　转换技术 …… 33

2.3.2　双栈方法 …… 33

2.3.3　打隧道（协议封装）方法 …… 36

2.3.4　转换方法 …… 44

2.4　路由 …… 50

2.4.1　网络架构 …… 51

2.4.2　路由核心知识 …… 52

2.4.3　路由协议 …… 55

2.5　多穴连接 …… 75

2.5.1　因特网结构 …… 76

2.5.2　主机多穴连接 …… 77

2.5.3　站点多穴连接 …… 82

2.6 移动性 …… 89
2.6.1 移动 IPv4 …… 90
2.6.2 移动 IPv6 …… 91
参考文献 …… 95
第3章 因特网云内部的路由 …… 100
3.1 网络、因特网和层 …… 100
3.1.1 层相互作用 …… 103
3.1.2 因特网基础设施（因特网云内部是什么） …… 103
3.2 网络和路由 …… 105
3.2.1 IP 寻址 …… 105
3.2.2 网络和流量：电路和分组（数据报）交换 …… 108
3.2.3 网络设备 …… 109
3.2.4 网络流量路由 …… 111
3.3 路由协议 …… 123
3.4 主要路由协议 …… 128
3.4.1 路由信息协议 …… 128
3.4.2 内部网关路由协议 …… 132
3.4.3 增强的内部网关路由协议 …… 134
3.4.4 开放最短路径优先 …… 140
3.4.5 路由器类型 …… 145
3.4.6 边界网关协议 …… 146
3.5 组播路由 …… 149
3.5.1 组播编址指派 …… 151
3.5.2 组播组 …… 153
3.5.3 组播树 …… 154
3.5.4 组播转发 …… 157
3.5.5 组播路由算法 …… 158
3.5.6 组播组成员关系协议 …… 160
3.5.7 组播路由协议 …… 161
3.6 虚拟路由器和负载均衡 …… 165
3.7 基于策略的路由 …… 167
3.7.1 引言 …… 167
3.7.2 策略路由 …… 168
3.7.3 策略路由结构 …… 168
3.7.4 实现策略路由 …… 169
3.8 路由器和交换机：平台架构 …… 171

3.9 安全管理 ······ 176
3.9.1 OSPF ······ 177
3.9.2 BGP ······ 178
3.10 电信和公众网络：交换和路由 ······ 178
3.11 无线、移动、自组织和传感器网络中的路由 ······ 181
3.12 网络、复杂性的本质和其他创新 ······ 182
参考文献 ······ 185
第4章 全 IP 网络：移动性和安全性 ······ 188
4.1 引言 ······ 188
4.2 移动 IP ······ 189
4.2.1 发现 ······ 191
4.2.2 注册 ······ 191
4.2.3 打隧道 ······ 192
4.3 IPv6 的移动 IP ······ 192
4.3.1 移动 IPv6 的基本操作 ······ 193
4.3.2 移动 IPv4 和移动 IPv6 之间的差异 ······ 193
4.3.3 移动 IPv6 安全 ······ 194
4.3.4 移动 IPv6 中的切换 ······ 194
4.3.5 3G CDMA 网络之上移动 IPv6 中的切换 ······ 195
4.4 IP 网络中的安全 ······ 196
4.4.1 IPsec 如何工作 ······ 198
4.4.2 IPsec 中的各元素 ······ 199
4.4.3 外发 IP 流量处理（保护到未保护） ······ 201
4.4.4 进入 IP 流量处理（未保护到保护） ······ 201
4.5 融合网络中的认证、授权和计费 ······ 202
4.5.1 Diameter ······ 203
4.5.2 移动 IPv6 中的 AAA ······ 203
4.5.3 一个融合的移动环境的安全框架 ······ 204
4.5.4 3GPP 安全 ······ 204
参考文献 ······ 206
第5章 转换扩展的家庭：步向基于 IP 的异构以用户为中心融合环境的下一步骤 ······ 211
5.1 引言 ······ 211
5.2 新的全 IP 家庭场景 ······ 213
5.2.1 高清多媒体服务蓬勃发展 ······ 213

5.2.2 通信流的重新分发 …… 215
5.2.3 IP 家庭中的服务重新分发 …… 216
5.2.4 全 IP 家庭骨干的容量 …… 218
5.3 家庭（全 IP）骨干 …… 219
5.3.1 IP 作为家庭骨干网络的关键实体 …… 219
5.3.2 家庭网络相关的联网技术 …… 220
5.3.3 联网技术总结 …… 223
5.4 家庭网关 …… 224
5.5 桥接各项技术：步向全 IP 基础设施 …… 225
5.5.1 桥接全 IP 融合架构 …… 225
5.5.2 不需新导线作为全 IP 基础设施的一种解决方案 …… 229
5.5.3 物理媒介和协议融合 …… 230
5.6 全 IP 家庭网络基础设施之上的服务 …… 231
5.6.1 扩展家庭之上的随身使用四重播放服务 …… 231
5.6.2 e-健康应用 …… 232
5.6.3 隐私、安全和用户概要 …… 233
5.7 扩展家庭网络 …… 234
5.7.1 全 IP 融合网络上的垂直和水平传输 …… 234
5.7.2 全 IP 扩展家庭基础设施中的 QoS …… 235
参考文献 …… 236
第 6 章 无线车载网：架构、协议和标准 …… 240
6.1 引言 …… 240
6.2 实施主动安全 …… 241
6.3 车辆网络架构 …… 242
6.3.1 智能车辆 …… 242
6.3.2 路侧单元和车载单元 …… 243
6.3.3 车辆通信 …… 243
6.4 车辆应用 …… 246
6.4.1 安全相关的应用 …… 247
6.4.2 非安全（便利性、舒适度）应用 …… 248
6.5 车载标准 …… 250
6.5.1 陆地移动的通信接入 …… 250
6.5.2 汽车到汽车（汽车间）通信联盟 …… 253
6.5.3 车辆环境中的无线接入 …… 259
6.6 无线车载网络中的挑战 …… 261
6.7 小结 …… 262

参考文献 …… 263
第7章 下一代IPv6网络安全：步向自治的和智能的网络 …… 266
7.1 引言 …… 266
7.1.1 背景 …… 266
7.1.2 下一代IPv6网络 …… 269
7.1.3 本章结构 …… 272
7.2 相关工作、工具和协议 …… 272
7.2.1 入侵检测/防御系统概述 …… 272
7.2.2 监测网络流量 …… 275
7.2.3 分组采样和流采样 …… 277
7.2.4 深度分组检测 …… 280
7.3 IPv6网络安全和用户剖析的智能 …… 281
7.3.1 分析器 …… 282
7.3.2 中心式服务器 …… 283
7.4 小结 …… 284
参考文献 …… 285
第8章 物联网 …… 289
8.1 物联网：新型因特网 …… 289
8.1.1 引言 …… 289
8.1.2 社会影响 …… 290
8.2 物联网的特点 …… 290
8.2.1 典型的LoWPAN节点的特点 …… 290
8.2.2 LoWPAN …… 291
8.3 实现物联网的标准 …… 293
8.4 用于物联网的协议层 …… 295
8.5 用于物联网的IEEE 802.15.4——PHY和MAC …… 295
8.5.1 868/915MHz频带 …… 296
8.5.2 2.45GHz ISM频带 …… 296
8.5.3 网络拓扑 …… 297
8.6 IPv6 …… 298
8.7 6LoWPAN：在无线个域网之上传输IPv6 …… 299
8.7.1 LoWPAN帧格式和交付 …… 299
8.7.2 一个6LoWPAN的邻居发现 …… 302
8.7.3 6LoWPAN中的IPv6地址自动配置 …… 303
8.7.4 首部压缩 …… 304

8.7.5 6LoWPAN 网状路由 …… 310
8.7.6 LoWPAN 广播 …… 311
8.8 传输层 …… 311
8.9 应用层协议 …… 312
8.10 物联网的网络架构 …… 313
8.10.1 自治 LoWPAN …… 313
8.10.2 具有扩展因特网连接能力的 LoWPAN …… 313
8.10.3 真正的物联网 …… 314
8.11 安全考虑 …… 314
8.12 物联网的应用 …… 315
8.12.1 智能电网 …… 315
8.12.2 工业监测 …… 315
8.12.3 结构监测 …… 316
8.12.4 保健 …… 316
8.12.5 连接的家庭 …… 316
8.12.6 远程测量 …… 317
8.12.7 农业监测 …… 317
参考文献 …… 317
第9章 6LoWPAN：采用 IPv6 互联物体 …… 319
9.1 引言 …… 319
9.2 传感器节点 …… 320
9.3 IEEE 802.15.4 标准 …… 322
9.4 6LoWPAN …… 324
9.4.1 6LoWPAN 适配层 …… 324
9.4.2 6LoWPAN 路由 …… 325
9.4.3 网状网之下路由 …… 326
9.4.4 路由之上路由 …… 327
9.4.5 6LoWPAN 地址指派 …… 328
9.4.6 6LoWPAN 首部压缩 …… 329
9.4.7 6LoWPAN 分片 …… 330
9.4.8 6LoWPAN 邻居发现 …… 330
9.5 6LoWPAN 实现 …… 332
9.5.1 TinyOS …… 332
9.5.2 ContikiOS …… 333
9.6 小结 …… 334
致谢 …… 334

参考文献……………………………………………………………………………………… 335
第 10 章 光纤上的 IP ……………………………………………………………………… 337
10.1 引言………………………………………………………………………………… 337
10.2 封装中的网络数据…………………………………………………………………… 338
10.3 为什么需要帧………………………………………………………………………… 341
10.4 IP 和光网络 ………………………………………………………………………… 344
10.5 WDM 网络中的控制 ………………………………………………………………… 348
10.6 IP 域中的分组汇聚 ………………………………………………………………… 351
10.7 全 IP 光突发交换网络 ……………………………………………………………… 355
10.8 小结………………………………………………………………………………… 358
参考文献……………………………………………………………………………………… 359
第 11 章 WiMAX 上的 IPv6 ……………………………………………………………… 364
11.1 引言………………………………………………………………………………… 364
11.2 WiMAX 技术概述 …………………………………………………………………… 365
11.2.1 物理层………………………………………………………………………… 365
11.2.2 MAC 层 ……………………………………………………………………… 367
11.3 WiMAX 网络架构 …………………………………………………………………… 369
11.4 IPv6 和 WiMAX ……………………………………………………………………… 373
11.4.1 邻居发现……………………………………………………………………… 374
11.4.2 无状态自动配置 ……………………………………………………………… 375
11.4.3 WiMAX 和自动配置 ………………………………………………………… 376
11.5 在 WiMAX 上部署 IPv6 的挑战……………………………………………………… 376
11.5.1 组播支持……………………………………………………………………… 376
11.5.2 子网或链路模型 ……………………………………………………………… 377
11.6 对建议解决方案的讨论……………………………………………………………… 380
11.6.1 组播支持……………………………………………………………………… 380
11.6.2 BS 和 AR/ASN-GW 接口 …………………………………………………… 381
11.6.3 AR/ASN-GW 和 NDP 规程…………………………………………………… 383
11.6.4 子网模型……………………………………………………………………… 384
11.6.5 移动性………………………………………………………………………… 384
参考文献……………………………………………………………………………………… 385

第1章 全IP网络：导论

Asoke K. Talukder，Nuno M. Garcia，Jayateertha G. M.

术语“互联网”的诞生可追溯到1969年，当时因特网工程任务组（IETF）发行了第一个请求评述（RFC 1），这是一份可公开得到的文档，总结了因特网共同体在一个特定专题方面的贡献。这也被称作第一代因特网的诞生——研究人员的数据通信协议。RFC 1标题为Host Software（主机软件），讨论接口消息处理器（IMP）和主机到主机协议。IMP是分组交换节点，从20世纪60年代晚期到1989年，用来将参与的网络互联到高等研究项目署网络（ARPANET）。ARPANET的官方名字是ARPA网络。ARPANET是在美国国防部（DoD）建立的，鼓励和资助高等科学和工程研究，在科学和技术方面将美国树立为领先者。在大约50年后，以及40多年以上的演进之后，因特网成为最具颠覆性技术之一，触及全世界每个人的生活，从婴儿到老人、富人到穷人，以及女人到男人。现在它是数据通信的主要载体，不管在简单电子邮件通信、社交网络的语境中，还是组织大众运动的一个工具方面都是如此。

1.1 各代因特网

IMP是第一代网关，如今被称作路由器。这是数据网络互联网基础建立之处（根基）。Ray Tomlinson，当时作为Bolt Beranek和Newman（BBN）技术公司的一名计算机工程师，在1971年底发明了基于因特网的电子邮件，这成为因特网中最流行的应用之一。在1971年通过RFC 113引入了文件传输协议（FTP）。Telnet规范RFC 137也在1971年得以发行。在接下来的一年，即1972年中，通过RFC 360，引入了远程任务入口（RJE），它集成了Telnet和FTP。也是在1972年10月，Larry Roberts和Robert Kahn在美国华盛顿特区举办的国际计算机通信会议（ICCC）上展示了ARPANET。1973年春天，Vinton Cerf，现存ARPANET网络控制程序（NCP）协议的开发人员，加入Kahn，研究开放架构互联模型，目标是为ARPANET设计下一代协议。之后ARPA与BBN技术公司、斯坦福大学和英国伦敦大学学院（University College at London）签订合同，在不同硬件平台上开发通信协议的可运行版本。开发了传输控制协议的四个版本：TCPv1，TCPv2，在1978年春天分隔成TCPv3和IPv3，之后稳定为TCP/IPv4——仍然在因特网上使用的标准协议[1]。这个协议在1980年通过RFC 760进行发布。

有些研究人员建议1971年为因特网的诞生年。事实上，可以说，带有小写字

母 i 的 internet（互联网）是 1969 年诞生的，带有大写字母 I 的 Internet（因特网）是 1971 年诞生的。但是，在 1980 年随着在因特网协议（IP）栈中引入 IPv4，因特网达到其成年阶段。互联网仅包含用于数据网络的 TCP 和 IP 通信协议，与此不同的是，因特网涵盖网络和应用协议栈，包括电子邮件、Telnet、FTP 和 RJE。不管互联网还是因特网，这一代分组交换网络和在这些网络上的应用都仅为研究共同体使用。相反，在那些日子里，业界正在使用其自己的专有协议集，如 IBM 的系统网络架构（SNA）和 DEC 的 DECnet，包括由应用和数据通信协议组成的一个产品套件。

很快因特网发生演进，可称这为第二代（2G）因特网。2G因特网是通用数据通信协议。可回溯到 1989 年，当时以规范的方式将域间路由包括到因特网中，这些规范如开放最短路径优先（OSPF）协议（RFC 1131）、边界网关协议（BGP）（RFC 1105）和 IP 组播（RFC 1112）。这些网络协议有助于在美国外的其他地区接受因特网，它成为一个网络互联（Interworking）协议，并将其影响扩散到世界各地。澳大利亚、德国、以色列、意大利、日本、墨西哥、荷兰、新西兰和英国都加入因特网[2]。主机数量从 1989 年 1 月的 80000 翻倍到 1989 年 11 月的 160000 多台。

之后随着 Tim Berners-Lee 发明超文本传输协议（HTTP），出现了第三代（3G）因特网。他是在 1990 年编写第一个万维网（Web）客户端和服务器的。他的统一资源标识符（URI）（通过 RFC 1738 定义）、超文本标记语言（HTML）（RFC 1942）和 HTTP（RFC 1945）等规范，帮助常人将因特网用于通用信息访问。很快话音在 1995 年被集成到因特网之中，是通过 IP 上的话音（VoIP）协议做到这一点的。这两项开创性的技术在因特网中的融合，帮助因特网走向成熟，并成为通信的一个通用媒介——不管通信是数据、信息、话音、图像，还是多媒体。在世纪之交，随着越来越多的服务被集成到因特网，因特网的领域开始拓展。RFC 数据库从 RFC 1 到达 RFC 1945，用了 27 年，但 HTTP 发行之后，在仅仅 13 年中，就有 4000 个以上的 RFC 被添加到因特网规范族（Suite）。从这个演进看，下一代因特网（NGI）的出现是明显的。NGI 将克服因特网的多数当前不足，这使之成为通用通信和服务的技术平台。

1.2 无线因特网

来自于受限于一个确定空间的自由（要求），是使研究人员和业界希望将无线通信集成到目标技术的一项强大驱动要素。因特网也不例外——研究界和业界进行了持续不断的工作，通过创新的无线技术使通信成为移动的。在过去十年左右的时间内，因特网使用的指数增长（在前面篇幅中可追溯其演进过程），对无线通信具有巨大影响。无线因特网在技术演进和开发的意义下一直在发生突发增长，从提供

话音服务到提供数据服务，当前正走向在线实时多媒体连接能力（Connectivity）。无线网络市场的演进是可采用逻辑方式追溯的，将之分成三类：面向话音的市场、面向数据的市场和在线多媒体连接能力（的市场）。

面向话音的市场围绕无线连接演进到公众交换电话网络（PSTN）。这些服务进一步演进为局域和广域市场。局域面向话音的市场基于低功率、低移动的设备，具有较高的话音质量。局域面向话音的应用开始于无绳电话的引入（在20 世纪 70 年代），它使用自第二次世界大战以来就存在的步话机（Walkie-Talkies）中使用的类似技术。第一部数字无绳电话使用在 20 世纪 80 年代早期在英国开发的 CT-2 标准。之后第二代无绳电话是无线专用局交换机（Private Branch eXchange，PBX），它使用数字欧洲无绳电话（DECT）标准。在简单的无绳电话之外和在一个较大区域及多应用之上，CT-2 和 DECT 都有最少的网络基础设施。这些局部系统很快演进成个人通信系统（PCS），这是带有其自己的基础设施的一个完备系统，非常类似于蜂窝电话网络。但是，总之，没有哪个 PCS 标准取得商业上的成功，所以，在 20 世纪90 年代后期，与蜂窝电话工业合并，取得一项巨大的商业成功。蜂窝网络的思想是非常老的：在 1947 年，AT&T 贝尔实验室提出了频率重用的思想，方法是将覆盖面积分成较小的蜂窝。但是，由于各种许可证和商业问题，蜂窝移动电话技术在当时没有发展腾飞[3,4]。

广域面向话音的市场，围绕蜂窝移动电话服务进行演进，这些服务使用具有高功率消耗、深度覆盖（Comprehensive Coverage）和低话音质量的终端。第一代无线移动通信是基于模拟信令的。在美国实现的模拟系统被称作模拟移动电话系统（AMPS），而在欧洲和世界其他地区实现的系统被称之为全接入通信系统（TACS）。模拟系统主要基于电路交换技术，是单为话音而不是数据进行设计的。2G 移动网络基于低频带（Low-Band）数字数据信令。最流行的 2G 无线技术被称作全球移动通信系统（GSM）。GSM[5] 首次是在 1991 年实施的。GSM 技术是频分多址/时分多址（FDMA/TDMA）的组合体，现在运行在大约 140 个国家。一种类似的技术称作个人数字通信（PDC），使用 TDMA 技术，是在日本出现的。自那时之后，在世界各地部署了几个其他基于 TDMA 的系统。GSM 是在欧洲开发的，而码分多址（CDMA）是在美国开发的。CDMA 使用扩频技术将话音分成小的数字化分段。CDMA 技术被公认为，以较小的背景噪声、较低的丢弃呼叫、增强的安全性以及较高的可靠性和网络容量，提供比较清晰的话音。2G 系统是基于电路交换技术的。2G 无线网络是数字的，并将应用范围扩展到比较高级的话音服务。虽然 2G 无线技术可处理一些数据能力，如传真和短消息服务（SMS），但数据速率仅达到 9.6kbit/s。

无线面向数据的市场围绕因特网和计算机通信网络基础设施发生演进。无线面向数据的服务可被分成宽带、自组织和广域移动数据市场。无线局部网络为少量用户支持较高数据速率和自组织（Ad Hoc）操作。无线局部网络通常被称作

无线局域网（WLAN）。主要的WLAN标准是IEEE 802.11。最初是在1980年引入的，并用了几乎10年的时间才完成。自此之后，这项技术从IEEE 802.11演进到IEEE 802.11 a/b/g/e/n，成为一项强大的无线技术，支持的数据速率从2Mbit/s、11Mbit/s、54Mbit/s直到600Mbit/s。它工作在工业、科学和医疗（ISM）2.5GHz频带，并使用直接序列扩频（DSSS）、正交频分复用（OFDM）和多输入多输出（MIMO）技术。自组织网络包括无线个域网（WPAN），如蓝牙、红外和近场通信（NFC）。WPAN的覆盖要小于WLAN的覆盖，且它们被设计为支持个人设备，如笔记本电脑、蜂窝电话、头戴设备、话筒（Speaker）和打印机，在没有任何走线的情况下将它们连接在一起。蓝牙是用于自组织联网的技术，是在1998年引入的。像WLAN一样，蓝牙工作在ISM频带下，但工作在低数据速率，并使用面向话音的无线访问方法，这为话音和数据服务的集成提供了一个较好的环境。NFC或射频标签工作在仅有数厘米的非常近距离的邻域范围内[3]。

广域无线数据市场为移动用户提供互联网接入。属于这个类别的技术有2G+和全球微波接入互操作性（WiMAX）技术。GSM、PDC和其他基于TDMA的移动系统提供商和运营商开发了2G+技术，这是基于分组的，并将数据通信速度增加到高达384kbit/s。这些2G+系统基于如下技术：高速电路交换数据（HSCSD）、通用分组无线服务（GPRS）和增强的全球演进数据速率（EDGE）。HSCSD，一种电路交换技术，将数据速率改进到高达57.6kbit/s，方法是引入14.4kbit/s数据编码和汇聚了4个14.4kbit/s的无线信道时槽。GPRS是一个中间步骤，设计为支持GSM世界，在没有等待3G系统的全规模部署条件下，实现一个全范围的因特网服务。GPRS技术是基于分组的，并被设计为与2G GSM、PDC和TDMA系统（用于话音通信）并行工作，并从位置注册数据库得到GPRS用户概要。GPRS使用1个到8个无线信道时槽（采用200kHz频带分配）的整数倍作为一个载波频率，支持数据速度高达115kbit/s。数据被分组化，并使用一个IP骨干在公共陆地移动网络（PLMN）上传输，从而移动用户可访问因特网上的服务，如基于简单邮件传输协议（SMTP）/邮政协议（POP）的电子邮件以及基于FTP和HTTP的万维网服务。通过增加每个时槽的吞吐量[4]，EDGE标准改进GPRS和HSCSD的数据速率。

围绕高速因特网连接能力和实时多媒体通信，无线实时多媒体市场一直在演进着。在这个领域中的主要技术有3G、WiMAX和WiFi。3G技术代表从以话音为中心的服务到面向多媒体（话音、数据和视频）的转移。3G移动设备和服务已经将无线通信转变为在线实时连接能力，提供位置特定的服务，这可应需地提供信息。3G无线技术代表了各种2G无线通信系统融合为单一全球系统，这包括陆地和卫星部件（Component）。3G使用三种空中接口做到这一点：宽带CDMA、cdma2000［也称作国际移动电信（IMT）-多载波，或IMT-MC］和统一无线通信（UWC）-136。

通过这些技术，3G 系统提供良好的话音质量、较高的数据速率（使移动服务用户得到高速因特网）和多媒体连接能力。

WiMAX 技术是作为一项宽带无线通信标准开发的，以具有高速因特网连接能力，提供高速数据速率的无线服务。在数年内，它就成为宽带无线通信的事实标准，提供 3G 系统的强有力的竞争者。

WiMAX 为固定/移动用户提供无线宽带服务的泛在交付。当前的移动 WiMAX 技术主要基于 IEEE 802.16e 标准，该标准规范了一个正交频分多址（OFDMA）空中接口，并提供对移动性的支持。WiMAX 提供灵活的带宽分配、服务质量（QoS）多个内建类型的支持和高达 100Mbit/s 的标称数据速率（覆盖范围为 50km）。同样，WiMAX 为多媒体服务的部署做好了就绪准备，这些服务如 VoIP、视频点播（VoD）、视频会议、多媒体聊天和移动娱乐。

此外，可追溯到前面各段内容，具有高速连接能力的因特网用户数的指数增长（由于 IPv4 应用中的持续不断的发展，构建在 TCP/IP 协议族，高速路由器的工程化，以及内建的 QoS 机制）使 IP 成为传输和路由多媒体实时分组的无争议的标准。所以，从网络运营商角度看，多媒体访问无线技术 3G、WiMAX 和 WiFi 都利用 IP 作为传输和路由其分组的一种方法。考虑到无线业界的指数增长，以及接着在最近时候引入的各种无线设备，通过 IP 连接能力的这种无线和因特网技术的融合，为许多种设备（有线的和无线的）通过 IP 进行连接，打开了多种可能性。所以，在无线通信之上提供高速因特网连接能力，是正在逐步成熟技术融合的主要值得称赞的任务。

由此看到，下一代网络不仅是计算机的一个网络，而且是各种网络的一个连通混合体，这些网络有不同的物理层性质，有许多网元和设备，如个人计算机、笔记本电脑、便签本、移动设备和个人数字助理（PDA），使用各种应用（从话音和数据到具有移动性的实时多媒体通信）。

明显地，这个扩展和用途场景，很快暴露了 IPv4（就其地址空间而言）、QoS 处理、安全性和扩展性方面的限制。很快，也变得明显的是对这样一个 IP 的需要，即它不仅支持大规模路由和寻址，而且能够施加这种任务上的低开销（这项需求来自无线媒介通信领域），同时支持自动配置、内建认证和机密性以及与当前这代设备的互联，这包括移动性作为一个基本元素。

为解决这个问题，设计了 IPv6，来替换 IPv4。IPv6 将 IP 地址长度从 32 比特扩展到 128 比特，即 IPv6 支持 2^{128} 个地址，或约 3.4×10^{38} 个地址。这支持为 2010 年地球上每个人（估计有 68 亿人口）分配大约 5×10^{28} 个地址。此外，IPv6 被设计为处理因特网的增长速率，并以其内建的 QoS 和安全特征，处理对数据速率、服务、移动性和端到端安全性方面的急迫需求。

1.3 全IP网络

在本书中，包括了有关下一代网络的各章，该网络提供所有种类的多媒体服务，其中连接和通信是通过常见的网络层协议 IPv6 实现的。为方便读者考虑，在本书中的专题分为三组：联网、专用服务和高级通信。有关联网的专题，一般而言，讨论下一代网络的连接能力功能特征，像寻址、交换、路由、多穴连接、移动性和安全。有关专用服务的专题处理特定的网络服务和应用。有关高级通信的专题讨论特定物理层技术上的 IPv6。

1. 联网专题

在这个分类中包括三章，即 Jayateertha 和 Ashwini B. 撰写的“IPv6 中的寻址和路由”、Dattaram Miruke 撰写的“因特网云内部的路由”和 Asoke K. Talukder 撰写的“移动性和安全性”。

IPv6 中的寻址和路由：本章讨论 5 个专题——寻址、IPv4 到 IPv6 转换、路由、多穴连接和移动性。寻址一节深入讨论有关 IPv6 寻址的所有方面，像表示、分类和指派。IPv4 到 IPv6 转换一节主要讨论共存于 IPv4 架构内的三项主要过渡技术，并提供到一个纯 IPv6 基础设施的大事件型（Eventual）过渡。路由一节描述 IPv6 中的路由现象。在解释像路由器、路由算法和路由表等的路由基础之后，本节详细描述 IPv6 语境中的三个主要路由协议：RIPv2、OSPFv3 和 BGP-4。多穴连接一节描述作为 IPv6 关键功能特征的多穴连接，详细描述主机多穴连接和站点多穴连接的概念。最后，在没有深入到高层次讨论的情况下，移动性一节描述 IPv6 中移动性的基本操作。

因特网云内部的路由：本章专门讨论下一代网络中的交换和路由。以路由协议算法和数据结构的讨论开始描述。详细讨论路由协议，这包括无线网络和传感器网络中的组播路由、基于策略的路由、路由和交换。本章也处理路由器和交换平台架构。

移动性和安全性：本章讨论针对 IPv4 和 IPv6 的 IP 中的移动性，也讨论高级移动功能特征，像漫游、IPv6 中的切换、3G CDMA 网络中的切换、移动 IPv6 和移动 IPv6 中的安全。在 rgw 安全一节，本章描述 IPv6 内建的 IPsec 协议。本章也触及在网络层提供的 IPsec 服务。

2. 专用服务

在这个分类中包括三章，即 Jose Bilbao 和 Jgor Armendariz 撰写的“转换扩展的家庭”、Rola Naja 教授撰写的“无线车辆网：架构、协议和标准”和 Artur M. Arsénio、Diogo Teixeria 和 João Redol 撰写的“下一代 IPv6 网络安全：步向自治的和智能的网络”。

转换扩展的家庭：本章分析在将用户的基础设施适配到多媒体服务变革过程中

在当前家庭和扩展的家庭场景中刚出现的问题，重点突出一种全 IP 架构。在本章结尾处，讨论 IP 扩展的家庭架构，包括为新的 IP 服务采用最适合架构中所涉及的挑战。

无线车辆网：本章在未来宽带车辆网络的设计方面，形成某种深邃见解，这种网络能够适应变化的车辆流量条件和变化的移动性模式。本章也将焦点放在车辆网络标准、车辆应用和 QoS 机制，目标是改进安保信息的任务关键型的传播（Critical Dissemination）。

下一代 IPv6 网络安全：本章给出针对用户剖析（Profiling）和安全目的，在监测流量领域中的不同工作。它也为下一代 IPv6 网络提供一种解决方案，它使用选择性过滤技术，并与一种引擎流量深度分组检测（DPI）组合使用，识别客户最频繁使用的应用和协议。由此，使各 ISP 以一种可扩展的和智能的方式优化他们的网络，就是可能的。

3. 高级通信

在这个类别中包括四章，讨论不同物理层技术上的 IPv6，即 Syam Madanpalli 撰写的“物联网”、Gilberto G de Ameida、Joel Rodrigues 和 Lui’s M. L. Oliveira 撰写的“6LoWPAN：采用 IPv6 互联物体”、Nuno M. Garcia 撰写的“光纤上的 IP”与 Jayateertha G. M. 和 Ashwini B 撰写的“WiMAX 上的 IPv6”。

物联网：本章介绍“物联网”，即因特网之上的一种低功率无线个域网（LoWPAN）。本章描述它的网络架构、协议栈和应用。它也处理 LoWPAN 之上 IPv6 的传输，并描述在实现物联网方面对 IPv6 的要求。

6LoWPAN：本章讨论 LoWPAN 和设备、IEEE 802.15.4 标准、6LoWPAN 规范和适配层，包括建议的 6LoWPAN 邻居发现优化。

光纤上的 IP：本章描述与物理层之上 IP 有关的问题，并特别地描述光网络［实现波分复用（WDM）］之上 IP 的架构和控制。本章也从一个无感知（Agnostic）观点，讨论数据汇聚的概念，介绍了一种 IP 分组汇聚和转换器机器。最后，它描述全 IP 光网络的一种可能架构，是使用光突发交换实现的。

WiMAX 上的 IPv6：本章为在 WiMAX 上部署 IPv6 的可行性带来曙光，其中考虑了网络工作组（NWG）提出的解决方案模型和涉及的问题。部署中的主要障碍是由于这样的事实，即 WiMAX 技术是基于点到多点架构的，其中在两个静态站（SS）/移动站（MS）之间的媒介访问控制（MAC）层处没有授权直接通信，相反所有通信都是在基站（BS）开始和结束的，这种情况的影响是，在 WiMAX 中的 MAC 处不支持组播通信，而 IPv6 的无状态自动配置功能特征则要求 MAC 层的组播实现。

参考文献

1. Computer History Museum, http://www.computerhistory.org/internet_history/internet_history_80s.html.
2. Raj Jain, "Internet 3.0: ten problems with current internet architecture and solutions for the next generation," Military Communications Conference, Washington, DC, October 23–25, 2006, http://www1.cse.wustl.edu/~jain/papers/gina.htm.
3. Asoke K Talukder, Hasan Ahmed, and Roopa R Yavagal, Mobile Computing Technology, Applications and Service Creation (2nd Edition), McGraw-Hill, 2011.
4. Kaveh Pahlavan and Prashant Krishnamurthy, *Principles of Wireless Networks, a Unified Approach*, Prentice Hall of India, 2008.
5. GSM 05.05, GSM Technical Specification, Version 5.1.0: May 1996, www.etsi.org.

第2章 IPv6中的寻址和路由

Jayateertha G. M. , B. Ashwini

2.1 引言

当前，因特网协议版本4（IPv4）服务的计算机市场一直是因特网增长的驱动要素。它构成当前因特网和不可计数的其他较小型的互联网。这个市场一直以指数速率在增长。用在因特网通信端点处的计算机，范围从个人计算机到超级计算机都有。多数计算机附接到局域网（LAN），且绝大多数是不移动的。

增长的下一阶段不可能单单由计算机市场驱动，而且该市场受到网络融合的影响：采用全IP网络的无线、有线、数据、话音和视频。这些市场将归为几个领域，且是极端大型的，还有一个崭新的需求集合，而这在IPv4部署的早期阶段是不明显的。因为话音、数据、视频和移动网络融合到IP网络成为了现实，许多个人计算设备随着其价格的降低和其能力的增加，看来一定会成为泛在的。这些计算设备的一项关键能力是，它们将被联网，并支持各种类型的网络附接（方法），如射频（RF）无线网络、红外附接和物理导线。所以，所有这些都要求网络互联技术，并需要一个共同的协议，该协议可在各种物理网络之上正常工作。

网络和数据的这种融合的另一项成果是，出现了一个联网的娱乐市场，即四重播放、视频点播等。随着数字高清电视世界的逼近，一台计算机和一台电视机之间的区别将消失。所以，对IP存在一项要求，即它不仅支持大规模路由和寻址，而且施加一项低的开销（源自无线媒介领域的需求），并支持自动配置、内建认证和机密性、与当前这代设备的互联和移动性等作为基本要素。人们设计了因特网协议版本6（IPv6），提供市场的扩展性、灵活性和要求，这是由于话音、数据、视频和移动性融合到IP网络，是对IPv4的一个革命性步骤。

所以，下面将焦点放在IPv6的这些设计功能特征上，包括寻址、路由、多穴连接和移动性，还有过渡性的技术挑战。

2.2 寻址

在世界范围内，因特网的快速增长在最近时间达到最远的地方，几乎耗尽了公开的4字节IPv4地址空间。所以，对扩展IP地址空间，就存在一个急迫的必要性。IPv6被看作下一代联网协议，被标准化以替代当前的IPv4，在RFC 2360中规

范，这可追溯到 20 世纪 90 年代中期，以解决快速消失的 IP 地址空间，当然还有其他方面。这不仅打开了 IPv6 的发展历程，而且促进了其他技术的发展，这些技术延长了 IPv4 地址空间的寿命预期。例如：

1）无类域间路由（CIDR），支持区域性的因特网注册机构（RIR）。

2）因特网服务提供商（ISP）地址空间的分配。

3）使用网络地址转换（NAT）技术的私有地址空间分配。

4）动态主机配置协议（DHCP）的开发，其能力是在按需基础上在许多用途间共享地址。

不管这些方案如何更好地利用 IPv4 地址空间，日渐增长的因特网用户基础由于不仅是话音、数据、视频和移动性方面的融合，而且有网络接口和网络设备方面的融合，这些导致在市场上出现各种各样支持 IP 的设备。这些设备接下来可增加 IPv4 地址空间消耗，所以减少了其可用容量[15,19]。

依据业界估计，在无线域，10 亿以上的蜂窝电话、个人数字助理（PDA）和其他无线设备将要求因特网接入，且每种设备都将需要其自己独特 IP 地址。IPv6 支持一个 128 比特的地址空间，并可潜在地支持大约 3.403×10^{38} 个独特的 IP 地址。采用这个大型地址空间方案，IPv6 具有为附接到因特网的每个和每种（所有）设备或节点提供独特地址的能力。

2.2.1 寻址概述

IPv6 将 IP 地址的尺寸从 32 比特增加到 128 比特。从而得到一个非常大型的 IP 地址池，这就可支持一个比较宽的寻址层次结构范围和多得多的可寻址节点。这消除了 IP 地址稀缺性，并由此消除了 NAT 部署。消除 NAT，得到一个简化的网络配置，并降低硬件/软件复杂性。大型的 IPv6 地址空间也非常适合联网家庭的未来愿景，其中各种仪器和物件都将在因特网之上被联网和管理。所以，从现在开始，无线和移动设备的部署将不会因为 IP 地址稀缺而受到阻碍了。

在本节开始时讨论 IPv6 地址表示，之后深入研究 IPv6 首部格式，它包含源和目的地的 IPv6 地址。也讨论 IPv6 地址的分类，并以一些特殊类型的 IPv6 地址讨论来结束本节。

1. 地址表示

一个 128 比特（16 字节）的 IPv6 地址表示为由冒号分隔的 8 个分量组成的序列。如下：

< comp. 0 >：< comp. 1 >：… < comp. 7 >

每个分量 <comp. i> 由 16 比特（0 或 1）组成，表示为 4 个十六进制数字。每个十六进制数字表示 4 比特，依据每个十六进制数字（0 ~ F）映射到其 4 比特二进制映射。如下：

0 = 0000 4 = 0100 8 = 1000 C = 1100

1 =0001 5 =0101 9 =1001 D =1101

2 =0010 6 =0110 A =1010 E =1110

3 =0011 7 =0111 B =1011 F =1111

注意，IPv6 地址中的十六进制字母不是大小写敏感的［请求评述（RFC）2373[2]］。下面是一些 IPv6 地址的例子：

4FDE：0000：0000：0002：0022：F376：FF38：AB3F

3FFE：80F0：0002：0000：0000：0010：0000：0000

2001：0660：3003：0002：0a00：20ff：fe18：964c

为比较简洁地表示一个 IPv6 地址，RFC 4291[12] 给出如下规则：

1）去掉任何 16 比特分量内的前导 0。

2）将零分量的任何连续组表示为一个双冒号，但仅能使用一次。

将这两条规则应用到上面的地址例子，它们可如下写出：

4FDE：：2：22：F376：FF38：AB3F

3FFE：80F0：2：：10：0：0 或 3FFE：80F0：2：0：0：10：：

2001：660：3003：2：a00：20ff：fe18：964c

注意，在一个 IPv6 地址表示中总是存在 8 个分量。

所以，计算带有单个双冒号的地址中有多少个零是容易的。但是，在有一个以上双冒号的地址中，就变为歧义的。

考虑一个 IPv6 地址，例如，4C62：0：0：56FA：0：0：0：B5。

可将这个地址缩写为 4C62：：56FA：0：0：0：B5 或 4C62：0：0：56FA：：B5，但不能是 4C62：：56FA：：B5，原因是不能对此无歧义地解码，因为它可表示为如下两个中的任何一个：4C62：0：0：0：56FA：0：0：B5、4C62：0：0：56FA：0：0：0：B5。

2. IPv6 首部格式

为得到一个 IPv6 地址（形成每个 IPv6 分组首部各字段之一）的直观感觉，需要理解 IPv6 分组首部的格式。所以下面简短地谈一下 IPv6 分组首部格式，解释其显著的特征[4]。

相比 IPv4，IPv6 有一个不同的分组首部结构。图 2.1 和图 2.2 做了最佳图示。如图所示，相比 IPv4，IPv6 分组首部做了简化。在首部后面，选项字段做了重构，且不再是 IPv6 分组首部的组成部分。

这使在中间节点处的 IPv6 分组首部处理变得更加容易。IPv4 首部的首部长度和总分组长度字段为净荷长度字段所取代。IPv4 首部中的［服务类型（ToS）］字段为流量类（TC）字段所取代。IPv4 首部中的存活时间（TTL）字段为 IPv6 首部中的跳限制字段所取代。IPv4 首部中的协议字段为 IPv6 首部中的下一首部字段所取代。最后，为提供服务质量（QoS），添加了一个新的流标记字段。



图 2.1 IPv4 分组首部格式

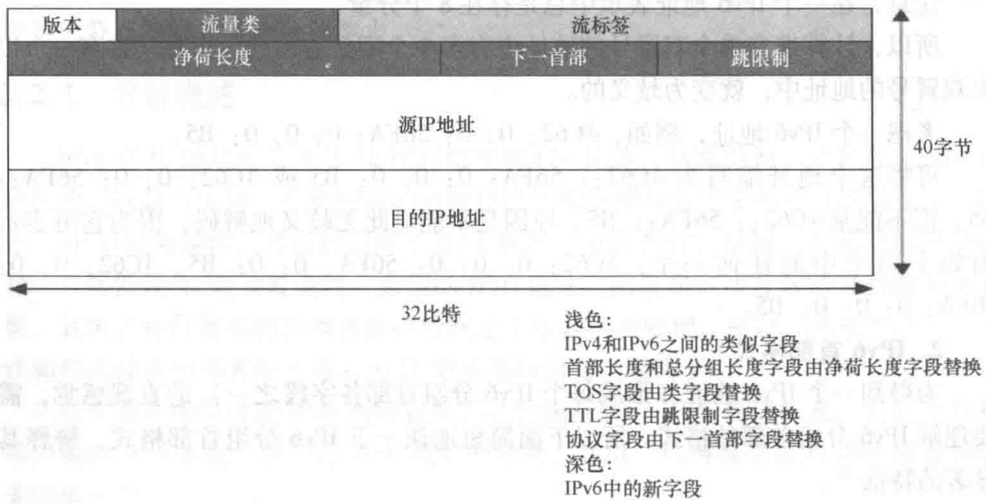


图 2.2 IPv6 分组首部格式

3. IPv6 地址前缀表示

一个IPv6 地址前缀表示，是 IPv6 地址前缀（或一个 IPv6 地址）及其前缀长度的组合表示。这采取如下形式：

IPv6 前缀（或 IPv6 地址）/前缀长度

其中 IPv6 前缀变量遵循 RFC 4291[2,12] 中规范的通用 IP 地址规则，前缀长度是一个十进制值，指明一个 IP 地址的连续高位阶比特的数量，这些比特形成 IPv6 地址的

网络部分。

注意 IPv6 前缀表示可被用来表示一个地址空间块（网络地址范围或网络），也可表示独特的 IPv6 地址。IPv6 地址前缀表示，用来表示一个网络地址，也称作无类域间路由表示法。通过考虑如下例子解释这一点：

［例 1］ 考虑 IP 地址 2001：CB8E：2A：：D15/64

扩展 IPv6 地址，得到：

2001：CB8E：002A：0000：0000：0000：0000：0D15

将 IPv6 地址扩展为完全的二进制格式，且因为前缀长度是 64，则可识别网络部分，方法是在 64 比特之后放上“/”，如下：

0010 0000 0000 0001 1100 1011 1000 1110 0000

0000 0010 1010 0000

0000 0000 0000/0000 0000 0000 0000 0000 0000

0000 0000 0000 0000

0000 0000 0000 1101 0001 0101

可将此格式写成 IPv6 前缀格式，如下表示该网络：

2001：CB8E：2A：：/64 是一个网络。

［例 2］ 考虑另一个例子：2002：3F0E：102A：：7/48。

2002：3F0E：102A：0000：0000：0000：0000：0007

（展开 IPv6 地址）

0010 0000 0000 0010 0011 1111 0000 1110 0001

0000 0010 1010/0000

0000 0000 0000 0000 0000 0000 0000 0000 0000

0000 0000 0000 0000

0000 0000 0000 0000 0000 0111（以完全的二进制格式进行全展开）

2002：3F0E：102A：：/48 是一个网络。

［例 3］ 考虑另一个例子：3FFE：10C2：43EE：D0C：F：：C15/126

3FFE：10C2：43EE：0D0C：000F：0000：0000：0C15

（展开 IPv6 地址）

也可如下展开并在 126 比特之后放上斜杠（/）来识别网络地址：

3FEE：10C2：43EE：0D0C：000F：：0000 1100 0001 01/01

3FEE：10C2：43EE：0D0C：F：：0C14/126 是一个网络。

从这些例子中看出，明显的是，网络前缀越小，则地址块就越大。

4. 地址类型

采用在联网中使用的主要寻址和路由方法，即单播寻址、任意播寻址和组播寻址，在参考文献中定义了 IPv6 地址的三个主类型。

一个单播地址识别单个独特的网络接口。IPv6 将发送到一个单播地址的一条

分组交付到特定接口。

一个任意播地址被指派到通常属于不同节点的一组接口。发送到一个任意播地址的一条分组，仅被交付到该组成员的一个成员，典型情况下是最近的主机，这是依据路由协议距离定义做出判断的。一个任意播地址与一个单播地址有相同格式，仅由其存在于网络中的多个点时才有区别。

一个组播地址为多台主机使用，通过参与到网络路由器间的组播分发协议，这些主机获得（Acquire）一个组播地址目的地。发送到一个组播地址的一条分组被交付到加入到组播组的所有接口。IPv6 没有实现广播地址。广播的传统角色被归入组播寻址到所有节点本地链路组播组。将在下面各节详细研究这些 IPv6 地址。

2.2.2 单播寻址

一个单播 IPv6 地址是识别一个 IPv6 接口的单一独特地址。ISP 将这些地址指派给组织。单播寻址提供全局独特地址。采用合适的单播路由拓扑，寻址到一个唯一地址的分组被交付到单个接口。

在 IPv6 中有几种类型的单播地址，即可聚合全局单播地址、本地用途地址、特殊地址、兼容地址和网络服务访问点（NSAP）地址。在未来可定义更多的地址类型。将在下面各节讨论这些单播地址。在开始讨论这些地址之前，先理解通用单播地址格式，这将帮助理解各种单播地址类型。

1. 单播地址格式

单播地址和任意播地址典型地由两个逻辑部分组成，用于路由的一个 64 比特网络前缀［全局网络前缀 + 子网标识符（子网 ID）］和用于识别主机之网络接口的一个 64 比特接口标识符（接口 ID）。网络前缀被包含在地址的最高 64 比特。RFC 6177[18] 建议，一个路由前缀的 56 比特分配给常规用户（如家庭网络），但一个 48 比特的路由前缀也是可能的。在这个场景中，8 比特子网 ID 字段可由网络管理员使用，定义给定网络内的子网。64 比特接口 ID 是自动从接口媒介访问控制（MAC）地址产生的，其中使用修改的扩展唯一标识符-64（EUI-64）格式［自动地从动态主机配置协议版本 6（DHCPv6）[10] 服务器得到的］，或手工指派的（见图 2.3）。

比特	56	8	64
字段	全局路由前缀	子网ID	接口ID

图 2.3 单播地址格式

2. 本地用途单播地址

有两种类型的本地用途单播地址，即链路本地用途地址［用于在线（on link）邻居之间］和站点本地用途地址（用于在同一站点中与其他节点通信的节点之间）。

（1）链路本地用途地址

这些基于一个接口 ID，它对网络前缀有一个典型格式（固定前缀 + 多个零），

如图2.4所示。它们用于到达附接到同一链路的邻居节点，且是由接口自配置的。所有IPv6地址有一个链路本地用途地址。

比特	10	54	64
字段	固定前缀	零	接口ID

图2.4 链路本地用途地址格式

前缀字段包含二进制值1111 1110 10。接下来的54个零使整个网络前缀为FF80::/64，对所有链路本地地址是同样的，这使它们为不可路由的。链路本地用途地址等价于自动私有IP寻址（APIPA）IPv4地址，它使用169.254.0.0/16前缀。链路本地用途地址的范围是本地链路。邻居发现协议（NDP）规程要求一个链路本地用途地址，且总是自动配置的。欲了解细节，可参见IPv6地址自动配置一节。

(2) 站点本地用途地址

站点本地用途地址也基于子网ID和接口ID，带有网络前缀的前48比特（10比特固定前缀+多个零），如图2.5所示。它们被指派到一个隔离内部网内的各接口。这可容易地迁移到一个基于提供商的地址，并等价于IPv4私有地址空间(10.0.0.0/8、172.16.0.0/12和192.168.0.0/16)。站点本地用途地址从其他站点是不可达的，且路由器不会将站点本地流量转发到站点外部。站点本地用途地址的范围是站点。

比特	10	38	16	64
字段	固定前缀	零	子网ID	接口ID

图2.5 站点本地用途地址格式

对于站点本地地址，前48比特总是固定的。固定前缀有二进制值1111 1110 11，与后跟的38个零，形成每个站点本地用途地址的前48比特：FEC0::/48。在这48个固定比特之后是一个16比特子网ID字段，采用该字段，可在站点内创建子网。在子网ID字段之后是64比特的接口ID，识别网络上的一个特定接口。除了地址的前48比特外，可聚合全局单播地址和站点本地用途的地址共享相同的结构。

3. 特殊的单播地址

在IPv6中存在具有特殊含义的一些单播地址。下面讨论这样的单播地址。

(1) 未指派地址

这是具有全零比特的一个地址，表示为0:0:0:0:0:0:0:0或::或::/128。它仅用于指明缺少一个地址，并等价于IPv4未指派地址0:0:0:0。未指派地址典型地用作这种分组的源地址，它们尝试验证临时单播地址的唯一性。未指派地址既不指派到任何接口，也不用作目的地址。

(2) 环回地址

这是一个单播本地主机地址，表示为0：0：0：0：0：0：0：1或：：1或：：1/128。它典型地被用来识别一个环回（虚拟）接口，所以，发送到这个地址的各分组被环回到相同主机或节点，从来就不会发送到任何接口。由此，这个地址支持一个节点/主机向自己发送分组，并等价于IPv4环回地址127.0.0.1。

4. 兼容单播地址

定义这些类型的单播地址，目的是辅助从IPv4到IPv6的过渡，并有利于IPv4和IPv6主机的共存。

(1) IPv4兼容的地址

这个地址持有一个内嵌的全局IPv4地址。这个地址具有格式0：0：0：0：0：0.w.x.y.z或：：w.x.y.z，其中w.x.y.z是一个公开IPv4地址的点分十进制表示，如：：129.144.52.38（在IPv6压缩格式中的相同地址将是：：8190：3426）。这些地址由双栈主机使用，以隧道方式在IPv4网络上传输IPv6分组。双栈主机是带有IPv4和IPv6栈的主机。当一个IPv4兼容的地址被用作一个IPv6目的地时，IPv6流量自动地被封装上一个IPv4首部，并在一个IPv4基础设施之上发送到目的地。

(2) IPv4映射的地址

这个地址也持有一个内嵌的全局IPv4地址。这个地址有格式0：0：0：0：0：FFFF.w.x.y.z或：：FFFF.w.x.y.z，其中w.x.y.z是一个公开IPv4地址的点分表示，如：：FFFF.129.144.52.38（以IPv6压缩格式表示为：：FFFF：8190：3426）。这个地址用来将公开IPv4主机的地址表示为IPv6应用的一个IPv6地址，这些应用使用AF_INET6套接字。采用这种方式，不管在IPv4还是在IPv6网络之上发生通信，这些IPv6应用总是以IPv6格式处理IP地址。重要的是指出，IPv4映射的地址仅用于内部表示。IPv4映射的地址从来不会被用作源地址或目的地址。IPv6不支持IP映射地址的使用。

(3) 6to4地址

这个地址由6to4隧道技术使用，识别6to4分组，并在IPv4网络上以隧道方式传输这些分组。欲了解有关6to4隧道技术的更多信息，请参见2.3节。6to4地址是这样形成的，即将前缀2002：：/16与主机的公开IPv4地址的32比特组合，形成一个48比特的前缀。例如，对于IPv4地址129.144.52.38，6to4地址前缀是2002：8190：3426：：/48。

(4) Teredo地址

这个地址用在Teredo隧道技术中[13]，该技术支持在IPv4网络之上IPv6分组的NAT穿越。一个Teredo地址具有如图2.6所示的格式。

Teredo前缀字段有值2001：：/32。标志字段指明NAT类型是全锥面（Full Cone）（值=0×8000）或受限的或端口受限的（值=0×0000）。客户端端口和客户端IPv4地址字段表示其相应值反转每个比特值的混杂值（Obfuscated Value）。

比特	32	32	16	16	32
字段	Teredo 前缀	Teredo服务器 IPv4地址	标志	客户端端口	客户端 IPv4地址

图 2.6 Teredo 地址格式

5. NSAP 单播地址

这个地址为将一个 NSAP 地址映射到一个 IPv6 地址提供了一种方法。一个 NAP 地址使用固定前缀 0000001，并将一个 IPv6 地址的 IPv6 比特的后 121 比特映射到一个 NSAP 地址。欲了解地址映射的细节，请参见参考文献［1，14］。

6. 可聚合的全局单播地址

定义在 RFC 2374[3] 中的这个地址，在 IPv6 网络上是全局可路由的和可达的。这个地址等价于一个公开的 IPv4 地址。所以，一个可聚合全局单播地址的范围是整个 IPv6 因特网。如名字所指明的，这些地址被设计为聚合的或归总（Summarized）的，产生一个高效的路由基础设施。这个地址由其格式前缀（FP）001 所识别。可聚合的全局单播地址格式如图 2.7 所示。其中，TLA ID 表示顶级聚合标识符；NLA ID 表示下一级聚合标识符；SLA ID 表示站点级聚合标识符。

比特	3	13	8	24	16	64
字段	FP	TLA ID	RES	NLA ID	SLA ID	接口ID

图 2.7 可聚合的全局单播地址格式

可聚合全局单播地址中的各字段可描述如下：

1）格式前缀字段，指明可聚合全局单播地址的 FP，且其值是 001。

2）顶级聚合标识符字段，指明单播地址的 TLA ID。TLA ID 识别路由层次结构中的最高层级。TLA ID 是由因特网指派号码权威机构（IANA）管理的，并被分配给各 RIR，RIR 接下来分配个体 TLA ID 给大型的全球 ISP。一个 13 比特字段，支持高达 8192 个不同的 TLA ID。在 IPv6 因特网路由层次结构最高层级的路由器被称作非默认路由器，原因是它们没有默认路由。事实上，它们仅以对应于所分配 TLA ID 的 16 比特前缀进行路由。

3）RES 字段为未来用途保留，其中扩展 TLA ID 或（NLA ID）字段的大小。

4）下一级聚合标识符字段，指明单播地址的 NLA ID。一个 24 比特字段被用来识别一个特定的客户站点。NLA ID 支持一个 ISP 创建寻址层次结构的多个层级，以便组织寻址和路由，并识别站点。一个 ISP 网络的结构，对非默认路由器是不可见的。

5）站点级聚合标识符字段，指明单播地址的 SLA ID。一个个体组织使用一个 16 比特 SLA ID 识别其站点内的子网。一个组织可使用这 16 个比特创建 65536 个子网或寻址层次结构的多个层级和一个高效的路由基础设施。一个客户网络的结构对

ISP是不可见的。

6）接口ID字段，即一个64比特字段，指明在一个特定子网上一个节点的接口。

聚合全局单播地址创建一个三层的拓扑结构，如图2.8所示。

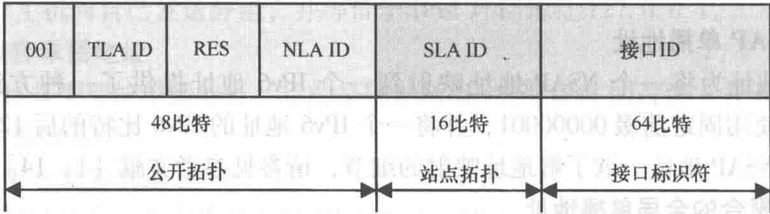


图2.8 全局单播地址格式的三层拓扑结构

公开拓扑是较大型和较小型ISP的一个集合体，它们提供到因特网的访问。站点拓扑是一个组织的站点内各子网的一个集合体。接口ID识别一个组织的站点内一个特定子网上的一个特定接口。

7. 唯一本地IPv6单播地址

唯一本地IPv6单播地址（ULA）是在RFC 4193[11]中定义的，拟用于通常在一个站点内部的本地通信。它有一个全局唯一的前缀，具有唯一性的概率，但并不期望在全球因特网上是可路由的。ULA在一个受限区域（如一个站点或站点的一个有限集合）内是可路由的。如此，它们是ISP无关的，并在没有任何永久的或间断性因特网连接的条件下，可用于一个站点内部的通信。有趣的是，即使它们通过路由或域名服务（DNS）被偶然地泄露到一个站点的外部，也与任何其他地址没有冲突。事实上，各应用可将这些地址看作全局范围的地址。否则，将前缀指派到这些地址可以是本地的，也可以是中心式指派的本地单播地址。唯一本地单播地址有如图2.9所示的格式。

比特	7	1	40	16	64
字段	前缀	L	全局ID	子网ID	接口ID

图2.9 唯一本地单播地址格式

前缀字段指明ULA，并有值FC00::/7。

L=1，如果前缀是本地指派的。

L=0，它可能在未来定义，但在实践中，用于中心式指派的前缀。

使用一个伪随机分配的全局ID，产生一个ULA，这意味着在各分配之间存在一个关系。所以，这些前缀并不打算用于全局路由。

8. 基于EUI-64地址的接口标识符

RFC 2373（4291）[2,12]陈述，使用从001到111前缀的所有单播地址，必须使

用从一个EUI-64地址派生得到的一个64比特接口ID。64比特EUI地址是由美国电气电子工程师学会（IEEE）定义的。EUI-64地址被指派到一个网络接口卡或是从网络接口卡的IEEE 802 MAC地址派生得到的。IEEE定义了从一个IEEE 802 MAC地址产生一个EUI-64地址的机制。一个EUI-64地址是与IEEE 1394（相线规范）兼容的，而且悄悄地取消了自动配置过程。

（1）IEEE 802 MAC地址

传统网络接口卡使用称作IEEE 802 MAC地址的一个48比特地址。这个地址由两部分组成：一个24比特的公司ID和一个24比特的扩展ID（也称作板ID），如图2.10所示。

比特	24	24
字段	ccccccug cccccccc cccccccc	xxxxxxxx xxxxxxxx xxxxxxxx
	IEEE管理的公司ID	公司选择的扩展ID

图2.10 IEEE MAC地址格式

在IEEE 802地址内一些特殊比特的含义如下：

1）全局/本地（U/L）比特是第一个字节的第7比特，并被用来确定该地址是全局管理的还是本地管理的。如果U/L=0，则这意味着该地址是由IEEE全局管理的。而如果U/L=1，则这意味着该地址是本地管理的。

2）组/个体（G/I）比特是第一个字节的最低比特，并被用来确定该地址是一个个体（组播）地址还是一个组（组播）地址。如果G/I=0，则该地址是单播，而如果G/I=1，则该地址是组播。

对于一个典型的802网络接口卡地址，U/L和G/I被设置为0，对应于一个全局管理的单播地址。

（2）IEEE 802 EUI-64地址

IEEE EUI-64地址表示网络接口卡寻址的一个新标准。它由一个24比特公司ID和一个40比特扩展ID组成，构成一个大得多的地址空间。EUI-64地址以与IEEE 802地址相同的方式使用U/L和G/L比特（见图2.11）。

比特	24	40
字段	ccccccug cccccccc cccccccc	xxxxxxxx xxxxxxxx xxxxxxxx xxxxxxxx xxxxxxxx
	IEEE管理的公司ID	公司选择的扩展ID

图2.11 EUI-64地址格式

（3）将一个IEEE 802地址映射到一个EUI-64地址

为从一个IEEE 802地址产生一个EUI-64地址，将16比特1111 1111 1111 1110（0xFFFE）插入到公司ID和扩展ID之间的IEEE 802地址。图2.12给出了一

个IEEE 802地址到一个IEEE EUI-64地址的转换过程。

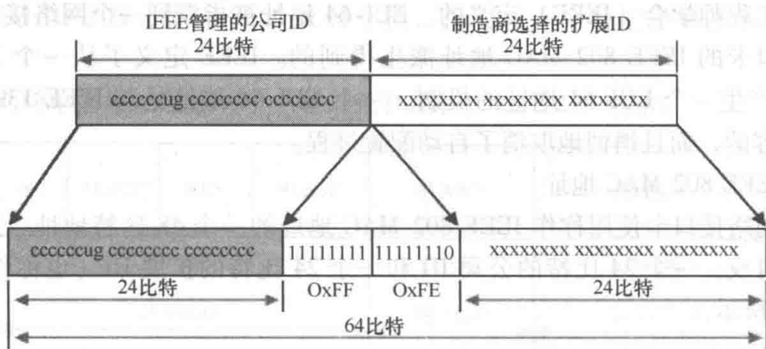


图2.12 映射的EUI-64格式

(4) 来自映射的EUI-64的IPv6接口标识符

为得到一个IPv6单播地址的一个64比特接口ID，在映射的EUI-64比特地址中的U/L比特取补（Complemented），即U/L比特被设置为1。图2.13给出了一个IPv6接口地址的全局管理的单播EUI-64比特地址。

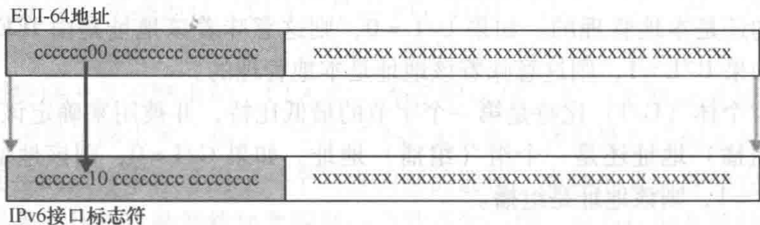


图2.13 IPv6接口ID

所以，为从IEEE 802地址得到IPv6接口ID，应该首先将IEEE 802地址映射到EUI-64比特地址，之后对U/L比特取补。图2.14给出了得到IPv6单播地址接口ID所涉及的所有步骤。

2.2.3 组播寻址

一个组播地址识别多个接口，且它用于一到多通信。如此，一个组播地址不能用作一个源地址。采用一个合适的组播路由拓扑，寻址到一个组播地址的各分组被交付到由该地址所识别的所有接口。一个组播地址由其FP 1111 1111（十六进制的FF）识别。它是依据几项特定的格式规则（取决于应用）形成的。一般而言，一个组播地址具有如图2.15所示的格式。

在一个组播地址中的各字段描述如下：

1）为任何组播地址，前缀字段持有二进制值1111 1111（十六进制的FF）。

2）标志字段指明要在组播地址上设置的标志。当前，定义了4个标志比特中

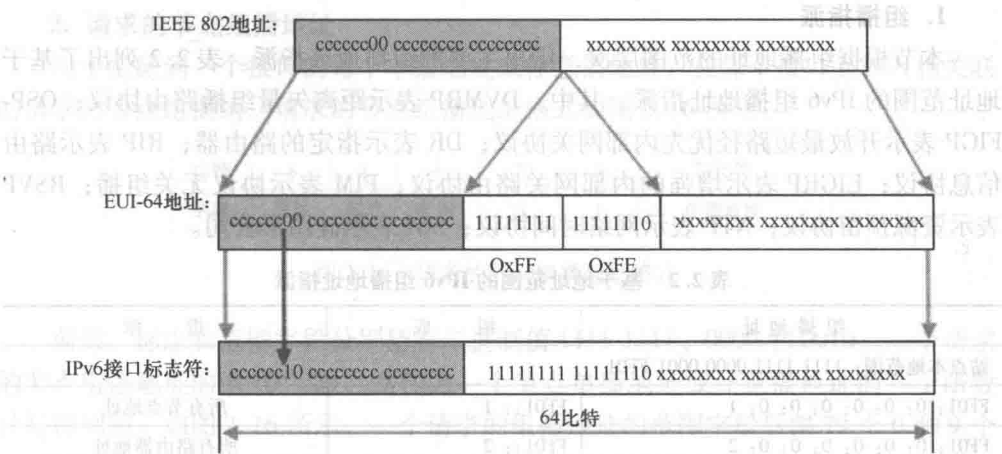


图 2.14　由一个 IEEE 802 地址得到的 IPv6 接口 ID

比特	8	4	4	112
字段	前缀	标志	范围	组ID

图 2.15　组播地址格式

的 3 个比特。最高位比特为未来用途保留。标志比特：O R P T，其中 O 为未来用途保留；T = 0，由 IANA 管理的永久地址；T = 1，临时组播地址；P = 1，由单播前缀派生得到；R = 1，内嵌的集结点地址。

3）范围字段指明组播所指的互联网络的范围。除了由组播路由协议提供的信息外，各路由器使用组播范围来确定是否要转发组播流量。表 2.1 给出了各范围及其相应的范围字段值。

表 2.1　IPv6 组播地址范围及其值

范围字段值	范　　围
1（0001）	接口/节点本地
2（0010）	链路本地
4（0100）	管理本地
5（0101）	站点本地
8（1000）	组织机构本地
E（1110）	全局

值 0、3 和 F 是保留的，而值 6、7、9、A、B、C 和 D 没有被指派。例如，带有组播地址 FF02::1 的流量具有一个链路本地范围。路由器从来不会将这种流量转发到本地链路之外。

4）组 ID 字段识别组播组，在范围内是唯一的。永久指派的组 ID 独立于范围。临时组 ID 仅与一个特定范围相关。

1. 组播指派

本节根据组播地址的范围层次，给出主要的组播地址指派。表 2.2 列出了基于地址范围的 IPv6 组播地址指派。其中，DVMRP 表示距离矢量组播路由协议；OSPFIGP 表示开放最短路径优先内部网关协议；DR 表示指定的路由器；RIP 表示路由信息协议；EIGRP 表示增强的内部网关路由协议；PIM 表示协议无关组播；RSVP 表示资源预留协议；NTP 表示网络时间协议；SGI 表示硅图形公司。

表 2.2　基于地址范围的 IPv6 组播地址指派

组播地址	组　成	指　派
站点本地范围：1111 1111 0000 0001 FF01		
FF01：0：0：0：0：0：0：1	FF01：：1	所有节点地址
FF01：0：0：0：0：0：0：2	FF01：：2	所有路由器地址
站点本地范围：1111 1111 0000 0005 FF05		
FF05：0：0：0：0：0：0：2	FF05：：2	所有路由器地址
FF05：0：0：0：0：0：0：3	FF05：：3	所有 DHCP 服务器
FF05：0：0：0：0：0：0：4	FF05：：4	所有 DHCP 中继
FF05：0：0：0：0：0：0：8	FF05：：8	服务定位
链路本地范围：1111 1111 0000 0002 FF02		
FF02：0：0：0：0：0：0：1	FF02：：1	所有节点地址
FF02：0：0：0：0：0：0：2	FF02：：2	所有路由器地址
FF02：0：0：0：0：0：0：3	FF02：：3	未指派
FF02：0：0：0：0：0：0：4	FF02：：4	DVMRP 路由器
FF02：0：0：0：0：0：0：5	FF02：：5	OSPFIGP
FF02：0：0：0：0：0：0：6	FF02：：6	OSPFIGP DR
FF02：0：0：0：0：0：0：7	FF02：：7	ST 路由器
FF02：0：0：0：0：0：0：8	FF02：：8	ST 主机
FF02：0：0：0：0：0：0：9	FF02：：9	RIP 路由器
FF02：0：0：0：0：0：0：A	FF02：：A	EIGRP 路由器
FF02：0：0：0：0：0：0：B	FF02：：B	移动代理
FF02：0：0：0：0：0：0：D	FF02：：D	所有 PIM 路由器
FF02：0：0：0：0：0：0：E	FF02：：E	RSVP 封装
FF02：0：0：0：0：0：1：1	FF02：：1：1	链路名称
FF02：0：0：0：0：0：1：2	FF02：：1：2	所有 DHCP 代理
FF02：0：0：0：0：0：FFXX：XXXX		请求的节点
全局范围：1111 1111 0000 0001 FF01		
FF0E：0：0：0：0：0：0：101	FF0E：：101	NTP 服务器
FF0E：0：0：0：0：0：0：102	FF0E：：102	SGI dogfight（混战）
FF0E：0：0：0：0：0：0：103	FF0E：：103	Rwhod

2. 请求的节点组播地址

对于指派到一个接口的每个单播地址或任意播地址，在那个接口上加入相关联的请求的节点组播组。请求的节点组播地址格式如图 2.16 所示。

比特	8	4	4	79	9	24比特
字段	前缀	标志	范围	0	1	单播地址

图 2.16　请求的节点组播地址格式

前缀、标志和范围字段分别持有二进制值 1111 1111、0000 和 0010。一个请求的节点组播地址的组 ID 字段，是作为一个节点单播地址或任意播地址的一个函数计算得到的。如图 2.16 所示，一个请求的组播地址的范围字段后跟 79 个 0 和 9 个 1，最后 24 比特是通过处理一个单播地址或一个任意播地址的最后 24 比特得到的。例如，对于带有一个链路本地单播地址 FE80::2AA:FF:FE28:9C5A 的一个节点，相应的请求节点组播地址是 FF02::1:FF28:9C5A。

请求节点组播地址有助于地址解析期间网络节点的高效查询。在 IPv4 地址解析协议（ARP）中，协议消息被发送到地址解析的 MAC 层广播。但在 IPv6 中，与通过使用一个本地链路范围所有节点地址而将信息分发到本地链路上所有 IPv6 节点的做法不同，请求节点组播地址被用作邻居请求（NS）消息的目的地。

2.2.4　任意播地址

一个 IPv6 任意播地址是典型地属于不同节点的一组接口的一个标识符。发送到一个任意播地址的一条分组，依据路由协议的距离度量，被交付到由那个地址识别的接口之一（距离最近的接口），如图 2.17 所示。它使用与一个单播地址相同的格式。所以，简单地通过检查地址，人们不能在一个单播地址和一个任意播地址之间做出区分。相反，任意播地址是以行政管理方式定义的。

任意播地址是从任意范围的单播地址空间得到的，不能从语法上与单播地址做出区分。任意播被描述为单播和组播之间的一项交叉功能（Cross）。像组播一样，多个节点可侦听一个任意播地址。像单播一样，发送到一个任意播地址的一条分组将被交付到那些节点之一（且只有一个）。由此，一个组播地址用于一到多通信，交付到多个接口，而一个任意播地址用于多个通信中的一到一通信，交付到单个接口。被交付到的准确节点是基于网络中的 IP 路由表的。

同样，为助于交付到最近的任意播组成员，路由基础设施必须知道被指派任意播地址的接口及其距离（就路由度量指标而言）。目前，任意播地址仅被用作目的地址，并仅被指派到路由器。在 RFC 2526[6] 和 RFC 4291[12] 中定义了保留的任意播地址。

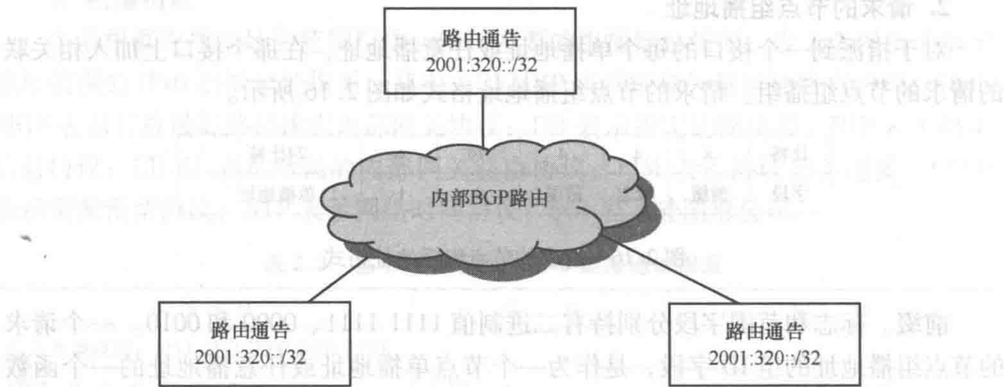


图 2.17 任意播的例子

最初，任意播寻址被称作集群寻址（Clustering Addressing）。这种寻址方式的动机来自于期望对服务复制的支持。例如，在一个网络上提供一项服务的一家公司，将一个任意播地址指派到几台计算机，这些计算机都提供该服务。当一名用户将一个数据报发往该任意播地址时，IPv6 将该数据报路由到集合（集群）中各计算机之一。如果另一名用户将一个数据报发送到一个任意播地址，IPv6 可选择将该数据报路由到集合中的一个不同成员，这使两台计算机可同时处理请求。

子网路由器任意播地址

对所有路由器而言，一个子网路由器任意播地址是预定义的和必备的。一个子网路由器任意播地址的格式如图 2.18 所示。

比特	N	128-*n*
字段	子网前缀	0

图 2.18 子网路由器任意播地址格式

它是从一个给定接口的子网前缀得到的。子网前缀中的各比特在其值方面是固定的，而剩下的比特都被设置为零。附接到一个子网的所有路由器接口，都被指派那个子网的子网路由器任意播地址。

2.2.5 主机和路由器的地址

在 IPv4 中，一般而言，带有单个网络接口的一台主机有单个 IPv4 地址，而有多个网络接口的一台路由器则有多个 IPv4 地址。IPv4 也以一个以上的 IPv4 地址来定义多穴连接主机。但是，在 IPv6 中，主机和路由器都有多个 IPv6 地址。例如，每台主机都有一个链路本地地址（用于本地链路流量）和一个可路由站点本地或全局地址。表 2.3 给出了一般情况下可向 IPv6 主机和 IPv6 路由器指派的地址。

表 2.3　主机和路由器的 IPv6 地址指派

地址类型	主　机	路　由　器
单播地址	每个接口一个链路本地地址	每个接口一个链路本地地址
	每个接口的单播地址，即一个站点本地地址或一个全局地址	每个接口的单播地址即一个站点本地地址或一个全局地址
	环回接口的一个环回地址	环回接口的一个环回地址
任意播地址	附加的任意播地址（手工方式或自动指派的）	每个子网一个子网路由器任意播地址
		附加的任意播地址
在组播地址上侦听流量	节点本地范围所有节点地址（FF01::1）	节点本地范围所有节点地址（FF01::1）
	链路本地范围所有节点地址（FF02::1）	节点本地范围所有路由器地址（FF01::2）
	在每个接口上每个单播地址的请求节点地址	链路本地范围所有节点地址（FF02::1）
	在每个接口上所加入组的组播地址	链路本地范围所有路由器地址（FF02::2）
		站点本地范围所有节点地址（FF05::2）
		在每个接口上每个单播地址的请求节点地址
		在每个接口上所加入组的组播地址

2.2.6　地址块分配

地址分配过程涉及在十六进制（IPv6 地址）和二进制之间来回转换，会涉及一个全局网络前缀，虽然有时会在一个十六进制数字上做出多数分配。但是，对于在做出分配之后，为管理剩下的块，二进制分解是必要的。在地址分配中使用的算法有三个基本类型。下面将列出这些类型。

1. 最佳拟合算法

这个算法基于这样的原则，即分配可满足所请求块尺寸需求的最小可用块。通过连续地将地址空间二分成所要求的尺寸[15]，这种方法高效地优化地址空间利用率。通过如下说明，解释这一点。

考虑带有一个全局网络前缀 4FFE：0320::/32 的一个网络。

假定希望为这个空间分配三个/34 网络。为做到这一点，首先将给定的网络表示以二进制表示，并连续地对之二分，如下：

0100 1111 1111 1110：0000 0011 0010 0000 ：0000 0000

0000 0000::/32

4FFE：320::/32

将这个网络二分成两个/33 网络，可通过如下分配做到：

0100 1111 1111 1110：0000 0011 0010 0000：1000 0000

0000 0000::/33

4FFE：320：8000::/33 和

0100 1111 1111 1110：0000 0011 0010 0000：0000 0000

0000 0000::/33

4FFE：320::/34

为进一步分割假定使用一个0值，1是空闲的（为以后分配）。可进一步将后者二分，如下：

0100 1111 1111 1110：0000 0011 0010 0000：0100 0000

0000 0000::/34

0100 1111 1111 1110：0000 0011 0010 0000：0000 0000

0000 0000 ::/34

4FFE：320::/34

由此，分配了两个块，以十六进制表示它们，则得到4FFE：0320：400::/34和4FFE：0320：0::/34。为得到第三个块，仍然如上那样将网络4FFE：320：8000::/33进一步分割，得到4FFE：320：C000::/34。

2. 稀有分配方法

在这种算法中，分配所要求尺寸的块，并简单地增加子网ID比特[15]。通过如下说明，解释这一点。

考虑带有一个全局网络前缀4FFE：320::/32的网络。假定需要分配大小为/40块的三个网络。如下分配：

0100 1111 1111 1110：0000 0011 0010 0000：0000 0000

0000 0000::/40

4FFE：320::/40

0100 1111 1111 1110：0000 0011 0010 0000：0000 0001

0000 0000::/40

4FFE：320：100::/40

0100 1111 1111 1110：0000 0011 0010 0000：0000 0010

0000 0000::/40

4FFE：320：200::/40

3. 随机分配方法

随机分配方法在子网络比特的尺寸范围内选择一个随机数，来分配子网络。以一个例子来说明这一点。考虑在前一节中一样的例子。给定带有一个全局网络前缀4FFE：320::/32的网络，分配大小为/40块的三个网络，在0和2^8-1（0~255）之间产生一个随机数，并实施分配，其中假定该数是存在的。这种方法为在所分配的实体间随机分布的分配提供一种方式，且一般来说，对于“相同尺寸”分配，

这种方法效果最佳。

2.2.7　单播或任意播地址指派规程

在 2.1 节 ~ 2.4 节详细讨论 IPv6 地址类型。注意，一个单播或任意播地址由两个逻辑部分组成：用于路由的一个 64 比特网络前缀和用来识别主机的网络接口的一个 64 比特接口 ID。本节将焦点放在这些 IPv6 地址是如何指派的或配置的。

一旦全局网络前缀由 IANA 指派，单播/任意播地址的 64 比特字段的最低比特，可以如下不同方式确定或指派：

1）从一个 48 比特 IEEE 802 MAC 地址自动配置得到，并扩展到一个 64 比特 EUI-64 格式。

2）在有状态地址自动配置期间通过 DHCPv6 指派得到。

3）依据 RFC 3041[7] 和 RFC 4941[17] 中的描述，为提供一定程度的匿名性，自动产生一个伪随机数，该数随时间改变。

4）手工配置的。

5）未来出现的其他可能方法。

由此，清楚的是，IPv6 主机可静态地指派地址（由管理员手工地输入），或动态地通过 DHCPv6（也称作有状态自动配置）得到，或通过自动配置（也称作无状态自动配置）和/或伪随机数产生得到。将在下面讨论这些技术。

1. 地址自动配置

地址自动配置是设备自动地配置其自己 IP 地址的一种能力，该地址对于它当前连接的子网是唯一的和相关的。存在 IPv6 地址自动配置的三种基本形式：

1）无状态的——不依赖于外部指派机制（如 DHCPv6）的状态或可用性。在没有外部或用户干预的条件下，该设备或主机尝试配置其自己的 IPv6 地址。

2）有状态的——单纯依赖于一个外部地址指派机制，如 DHCPv6。DHCPv6 服务器将指派 128 比特 IPv6 地址给主机，采取的是一种非常类似于 DHCPv4 操作的方式。

3）无状态和有状态的组合——涉及无状态地址自动分配（SLAAC）的一种形式，SLAAC 与带有附加参数的有状态配置一起使用。这常包括无状态地自动配置一个 IPv6 地址，但利用 DHCPv6 得到附加的参数或选项（如 DNS 来查询给定网络上的名字解析）。

将在下面讨论所有这些形式的自动配置。

（1）无状态自动配置

在无状态自动配置中，如前所述，在没有外部或用户干预的条件下，主机或设备尝试配置其自己的 IPv6 地址。下面给出在 RFC 2462（4862）[5,16] 中规定的无状态自动配置。通过实现如下步骤，可方便地做到：

1）一个接口 ID 的产生。IPv6 地址架构 RFC 2373（4291）[2,12] 规定，除了以

000 开始的那些地址外的所有单播 IPv6 地址，必须使用一个 64 比特接口 ID（使用修改的 EUI-64 算法）。欲了解细节，参见前面各节。

① 从 48 比特以太网 MAC 地址得到一个接口 ID 的算法：

• 逆转 MAC 地址公司 ID（初始 24 比特）字段的 u 比特。u 比特是公司 ID 字段中第 7 个最高位比特。

• 在公司 ID 和公司选择的扩展 ID（最后 24 比特）之间插入十六进制值 FFFE。

IPv6 地址的接口 ID 是逆转 u 比特的公司 ID 字段 +16 比特 EUI 标签 FFFE +24 比特扩展 ID。

• 例：令 AC-62-E8-49-5F-62 为主机的给定 MAC 地址。逆转 u 比特，得到 AE-62-E8-49-5F-62。插入 FFFE，得到一个 64 比特接口 ID，为 AE-62-E8-FFFE-49-5F-62。这可重写作 IPv6 接口 ID，为 AE62：E8FF：FE49：5F62。

② 对于非以太网 MAC 地址，该算法使用链路层地址作为接口 ID，从左开始填充零。

③ 对于以太网 MAC 地址，该算法使用链路层地址作为接口 ID，从左开始填充零。

④ 对于这样的情形，其中没有链路层地址可用。例如，在一条拨号链路上，建议利用设备上另一个接口地址的唯一标识符、设备的一个序列号或其他特定的标识符。

2）链路本地地址和重复地址检测。

① 依据 IPv6 地址架构，链路本地网络前缀是固定的。所以一旦产生一个接口 ID，主机可自动配置其链路本地地址，方法是将接口 ID 添加到一个预定义的 FE80::/64 前缀。利用前面的例子，它使用 MAC 地址 AC-62-E8-49-5F-62 和得到的接口 ID AE62：E8FF：FE49：5F62，则链路本地地址将是 FE80::AE62：E8FF：FE49：5F62。

② 重复地址检测是通过称作邻居发现的一个过程实施的，它包括这样的设备，它向刚派生得到的 IP 地址发送一条 IPv6 NS 分组，目标是识别 IP 地址的一个已存在的占有者。在一个微小的延迟之后，该设备也向与临时地址相关联的请求节点组播地址发送一条 NS 分组。

③ 如果另一个设备已经在使用该 IPv6 地址，它将以一个邻居通告（NA）分组做出响应，且自动配置过程将停止，即要求人工干预或配置设备来使用另外一个接口 ID。如果没有收到一个 NA 分组，则该设备假定地址的唯一性，并将之指派到相应的接口。

3）前缀发现，得到全局路由前缀和子网 ID。通过路由器通告（RA）消息，各路由器为所有本地连接的设备通告网络前缀和相关联的配置参数。一台主机或设备可发出一条路由器请求（RS）消息，显式地请求一条 RA 消息。这些 RA 消息可

由设备使用，以识别可用于这个子网的全局单播网络前缀。

4）网络前缀和接口 ID 的串接。设备或主机将其接口 ID 附加到网络前缀，并向这个临时地址发出 NS 分组。如果没有接收到 NA，则设备可将该地址指派到其相应的接口。

(2) 通过 DHCPv6 的有状态自动配置

用于 IPv6 的 DHCP 是由因特网工程任务组（IETF）通过 RFC 3315[10] 标准化的。DHCPv6 支持 DHCP 服务器传递配置参数（如 IPv6 地址）到 IPv6 节点。它提供可重用网络地址的自动化应用能力和附加的配置灵活性。这个协议是有状态的，并可与早期给定的无状态规程组合使用或单独使用，即独立于以前的规程加以使用。

如果一个客户端希望接收配置参数，它将在附接的本地网络上发出一条请求，以便检测可用的 DHCPv6 服务器。这是通过请求和通告消息完成的。著名的 DHCPv6 组播地址被用于这个过程。接下来，DHCPv6 客户端将从可用的服务器请求参数，它将以带有所请求信息的一条应答消息做出响应（见图 2.19）。



图 2.19　DHCPv6 规程

依据 RFC 2462（4862），现在描述 DHCPv6 中所涉及的实际步骤（见图 2.19）[5,16]：

1）客户端发出一条 SOLICIT 消息，在它所连接的子网上请求一个 IP 地址。这条 SOLICIT 消息被发送到 ALL_Relay_Agents_Servers 组播地址 FF02:: 1: 2。

2）配置为中继代理的任何路由器将接收 SOLICIT 分组，它将 SOLICIT 分组封装到一条 Relay_Forw 分组内，并将之转发到站点范围的 All_DHCP_Servers 组播地址 FF05:: 1: 3。

3）在相同子网上作为客户端的 DHCPv6 服务器，直接得到 SOLICIT 消息，而其他服务器通过 Relay_Forw 消息接收该消息。

4）DHCP 服务器将以一条通告分组做出响应，直接指明优先级值或通过中继代理得到，这取决于服务器是连接到客户端的子网上还是远程连接的。这个优先级值意图支持客户端选择带有最高优先级（由管理员配置的）的服务器。

5）客户端分析所接收到的通告，选择服务器，并发送单播请求消息，直接请求 IP 地址或通过中继代理得到。

6）DHCPv6服务器发送应答，确认所指派的IPv6地址（直接指派的或通过中继代理指派的）。

7）之后客户端实施重复地址检测规程，以便确定IPv6地址的唯一性。

（3）有状态和无状态自动配置的组合

在一台所附接路由器指令或默认网关（当存在时）的基础上，DHCPv6客户端将知道何时它希望使用DHCPv6。默认网关有在一条RA可用于这个目的的两个可配置比特：

1）O比特——当这个比特设置时，客户端可使用DHCPv6来检索其他配置参数（DNS地址）。

2）M比特——当这个比特设置时，客户端可使用DHCPv6从DHCPv6服务器检索一个受管理的IPv6地址。

当一台路由器发送带有O比特设置但没有设置M比特的一条RA时，客户端实施SLAAC，得到其IPv6地址，并使用DHCPv6得到另外的信息。另外信息的一个例子是DNS。人们都知道这种机制是无状态DHCPv6，原因是DHCPv6服务器不需要跟踪客户端地址绑定。

2. 通过随机接口标识符的地址自动产生

SLAAC过程使用一个全局唯一的和静态的IEEE 802 MAC地址，创建IPv6接口ID。这会导致用户设备由此对用户身份的跟踪，而不管在任何时间其网络前缀为何都可实施跟踪。为解决这个问题并提供对用户身份的一定程度的匿名性，RFC 3041（4941）[7,17]讨论基于时变随机字符串来产生一个IPv6接口ID的一种替代方法。基于随机接口ID得到的IPv6地址，被称作临时IPv6地址。注意，这些临时地址是针对公开网络前缀使用SLAAC产生的。这些临时IPv6地址可被用作产生连接的源地址。RFC 3041描述为IPv6系统（具有或没有存储能力）产生随机接口ID的规程。对于没有存储能力的IPv6系统，每次IPv6初始化时产生随机接口ID。对于具有存储能力的IPv6系统，历史信息被用来产生一个未来的随机接口ID。

对于具有存储能力的IPv6系统，在初始系统启动时，通过使用随机数，产生接口ID，并存储一个历史值。当下次IPv6初始化时，通过如下步骤产生一个新的接口ID：

1）从存储中检索历史值，并在网络接口卡的EUI-64地址的基础上，附加接口ID。

2）在步骤1的工作量只是计算消息摘要-5（MD-5）单向加密哈希（MD-5哈希计算超出了本章的范围）。

3）将在步骤2中计算的MD-5哈希的后64比特存储为下一次接口ID计算的历史值。

4）取在步骤2中计算得到的MD-5哈希的前64比特，并将第7比特设置为0。结果就是接口ID。

3. DNS 支持

对于成功的（地址协议）共存，需要一个 DNS 基础设施，原因是到处都存在名字的使用而不是地址的使用，来指代网络资源。升级 DNS 基础设施，由传播带有记录的 DNS 服务器组成，它支持 IPv6 名字到地址和地址到名字的解析。在使用一个 DNS 名字查询得到地址之后，发送节点必须选择为通信使用哪个地址。

（1）地址记录

为做到从域名到地址的成功解析，DNS 基础设施必须包含如下资源记录（手工或采用 DNS 动态更新实施传播）：

1）IPv4 节点的 A 记录。

2）IPv6 节点的 AAAA 记录。

（2）指针记录

为得到地址域名查询和反向查询的成功解析，DNS 基础设施必须包含手工或动态地传播的如下资源记录：

1）对于 IPv4 节点，IN_ADDR. ARPA 域的指针（PTR）记录。

2）对于 IPv6 节点，IPv6. ARPA 域的 PTR 记录。

2.3 IPv4 到 IPv6 转换

上一节讨论了 IPv6 如何解决“IP 地址耗尽”问题，方法是引入一个 128 比特 IP 地址结构，由此为下一代联网打开了一个海量的地址空间。不幸的是，在 IP 地址结构方面的这个改变，对网络栈具有一个重大影响。首先，当前 IP 层需要以一个新的 IP 层加以替换。且这个新的 IP 层需要调节以运行在如今普遍存在的各种 L2 机制上。那么当前传输协议［如 TCP、UDP 和其他新的传输机制（如流控传输控制（SCTP））］需要在新的 IP 层上重建。此外，如今构建于当前套接字层之上的多数应用，不得不使用新的套接字机制加以更新。

因为这些任务涉及大量的复杂性工作，所以抛弃现有的 IPv4 网络并立刻跳转采用 IPv6，根本是不可能的。所以，可预计，转换将以阶段式发生，首先一些 IPv6 节点被引入到 IPv4 网络，而这些节点的数量将随时间而逐渐增加，直到遥远的未来某个时刻，整个网络成为 IPv6 网络。本节将讨论迄今为止在文献中存在的各种转换技术，这使我们步向下一代网络，并最终完全采用 IPv6。为理解这个复杂的转换现象，需要理解 IP 网络架构中的变化，这在 RFC 2893[27] 中做了解释。RFC 2893[27] 定义了转换阶段期间如下的不同 IP 节点：

1）纯 IPv4 节点：仅实现 IPv4，且仅有 IPv4 地址，不支持 IPv6。如今安装的多数主机和路由器是纯 IPv4 节点。这示于图 2.20 的阶段 1 中。

2）纯 IPv6 节点：仅实现 IPv6，且仅有 IPv6 地址，不支持 IPv4。这个节点仅能够与 IPv6 节点和应用通信。这种类型的节点如今是不常见的，但随着较小型设

备（如蜂窝电话和手持计算设备）包括 IPv6 协议，这也许会变得更普遍。这示于图 2.20 的阶段 4 中。

3）IPv6/IPv4 节点：实现 IPv4 和 IPv6 的一个节点。见图 2.20 的阶段 2。

4）6to4 路由器：在其 IPv4 网络连接上配置有一个 IPv6 到 IPv4 伪接口的任何边界路由器。一台 6to4 路由器用作 6to4 隧道的一个端点（在 2.3.3 节详细解释），在这条隧道上，路由器将分组转发到另一个 IPv6 站点。

5）6to4 主机：一个接口配置有一个 IPv6 到 IPv4 派生地址的任意 IPv6 主机（在下面各节详细解释）。

6）站点：因特网私有拓扑的一部分，不为任何人和所有人携带中转流量。一个站点可跨越一个大型地理区域。例如，一个多国公司上的私有网络是一个站点。

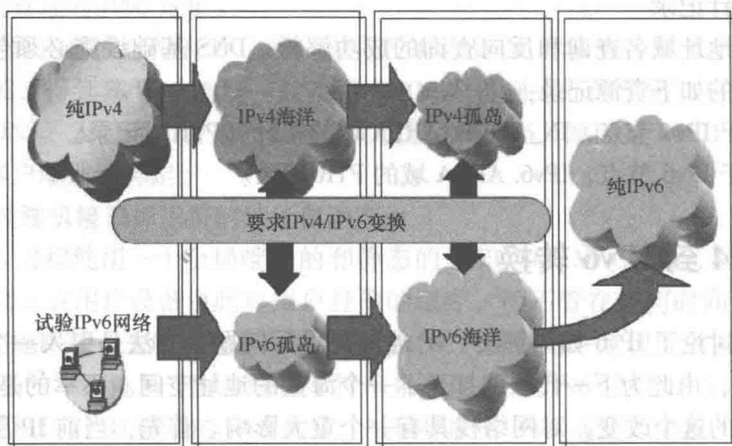


图 2.20　IPv4 到 IPv6 转换[31]

图 2.20 给出了在四个阶段间转换期间网络架构中改变的一个高层视图。注意阶段 1 由纯 IPv4 节点主导，带有拥有数个节点的一个试验型 IPv6 网络。

在第二阶段，纯 IPv4 节点的主导地位继续存在，但出现 IPv6 网络的合理增加。在第三个阶段，IPv6 网络占主导，同时 IPv4 网络孤岛仍然存在。最后阶段是全 IPv6 节点。同样注意到，在前三个阶段期间，要求使用 IPv4/IPv6 转换技术。

从到此为止的讨论中，非常明显的是，整个因特网仅运行在 IPv6 上是不可能的，原因是现有骨干 IPv4 网络会对 IPv6 的完全采用带来极大的困难。所以，在不久的将来，对两种联网技术的共存，存在强烈需要，且如此导致一组新的联网技术的出现。当所有 IPv4 节点都转换为纯 IPv6 节点时，才做到了真正的迁移。但是，在可预见的将来，取得实际的迁移，此时尽可能多的纯 IPv4 节点被转换为 IPv6/IPv4 节点。

2.3.1　转换技术

NG 转换工作组[27,32]已经定义了三种主要转换技术共存于 IPv4 基础设施内，并提供到一个纯 IPv6 基础设施的终极（eventual）转换。图 2.21 形象地给出了这些不同类型的转换技术。

1）双栈：在网络设备或主机上支持 IPv4 和 IPv6。

2）打隧道：将一条 IPv6 分组封装在一条 IPv4 分组内，以便在一个 IPv4 网络之上传输，也称作协议封装。

3）转换：地址的地址转换或端口转换，例如通过一个网关设备，或主机或路由器 TCP/IP 代码中的转换代码，也称作协议转换。

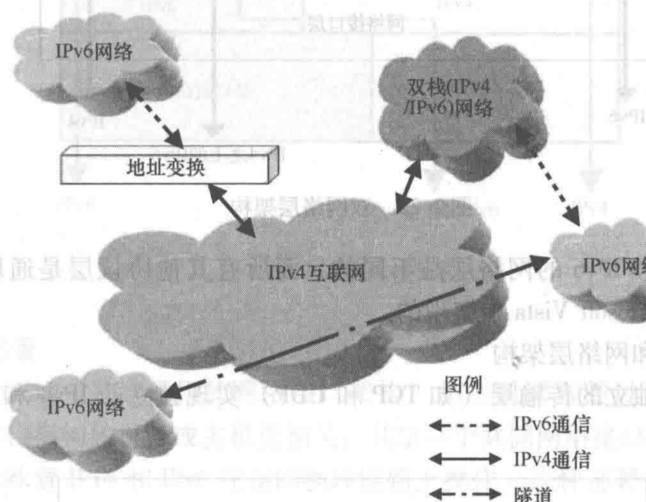


图 2.21　转换技术[31]

下面将详细研究这些转换技术。

2.3.2　双栈方法

这种方法要求在主机或设备（要求到两种联网技术的网络接入，包括路由器和网元）上实现 IPv4 和 IPv6 栈。这种设备需要配置有 IPv4 和 IPv6 地址，可通过为相应协议定义的方法得到这些地址。这使网络能够在转换时段期间支持 IPv4 和 IPv6 服务和应用。由此，这个时段接下来打开（拓展）了为 IPv6 服务出现的范围并变得可用。这项技术是容易使用的和灵活的。同样注意，双栈技术是其他转换机制的基础。例如，打隧道技术需要双栈端节点，而转换则要求双栈式的网关。

1. 双栈架构

依据功能需求和个体独特性，设备或主机、路由器及其他网元中的双栈实现，可能是不同的。这些实现方法可被粗分为如下三种架构[27,32]。

（1）双网络层结构

这包含 IPv4 和 IPv6 带有通用应用、传输和数据链路以及网络接口层实现的各层，如图 2.22 所示。

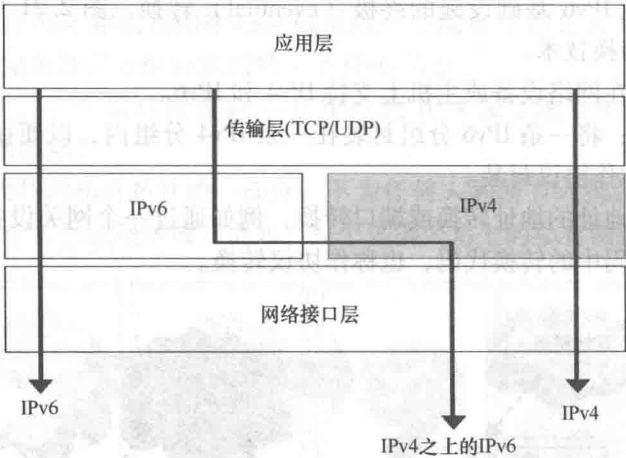


图 2.22 双网络层架构

注意，IPv4 和 IPv6 的网络层是不同的，而所有其他协议层是通用的。例如，这个架构是在 Microsoft Vista 中实现的。

（2）双传输和网络层架构

这种方法以独立的传输层（如 TCP 和 UDP）实现独立的 IPv4 和 IPv6 层，如图 2.23所示。

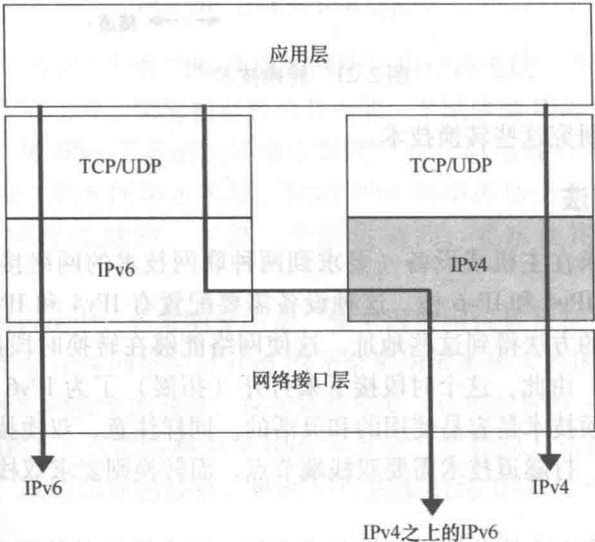


图 2.23 双传输和网络层架构

但这种方法有通用的应用、数据链路和网络接口层实现，如图 2.23 所示。例如，在 Microsoft XP 中采用这种方法。

（3）双整体栈架构

这种方法实现向下直到物理层实现的两个整体栈，对 IPv6 和 IPv4 要求独立的网络接口，如图 2.24 所示。这种方法可能是稀少的，但在有多项应用的网络服务器情形中期望采用这种方法，其中一些应用仅支持一个版本。

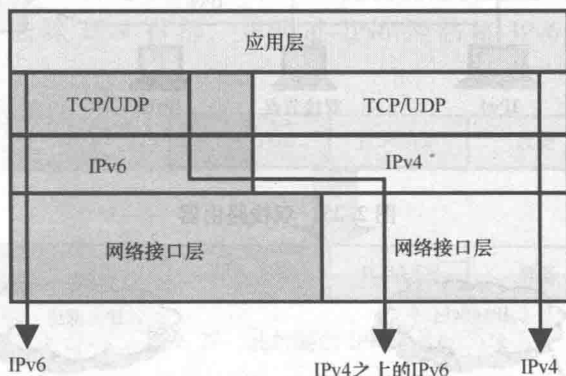


图 2.24　双整体栈架构

2. 双栈部署

在有一个双栈（共享通用的网络接口架构）的主机部署的过渡期间，实际上比具有双整体栈架构的设备或主机更相关。共享一个共同网络接口的双栈主机或设备的部署，意味着 IPv4 和 IPv6 在相同物理链路上操作。这种部署可从两个角度来看，即网络层角度和层 2 角度。

（1）网络层角度

这种方法利用如下事实，即层 2 技术（如以太网）支持 IPv4 和 IPv6 净荷。因为这样的主机和设备是带有共同 L2 层技术的双栈。同样，为在原生 IPv4 主机和支持 IPv6 主机间路由分组，各路由器也需要是双栈的。人们期望这种方法在转换期间是非常常见的，如图 2.25 所示。

（2）层 2 角度

RFC 4554[37] 描述使用虚拟局域网（VLAN）的一种新颖方法，它在不要求直接路由器升级条件下，支持双栈部署。这种方法要求 VLAN 标记（tagging），支持层 2 交换机将包含 IPv6 净荷的帧广播到一个或多个支持 IPv6 的路由器。一个 IPv6 VLAN 可配置为一台升级过的路由器（支持 IPv6）和一个以太网交换机端口（这些升级过的路由器接口连接到这里）。其他双栈设备可配置为这个 VLAN 的成员。图 2.26 形象地说明了这个部署场景。

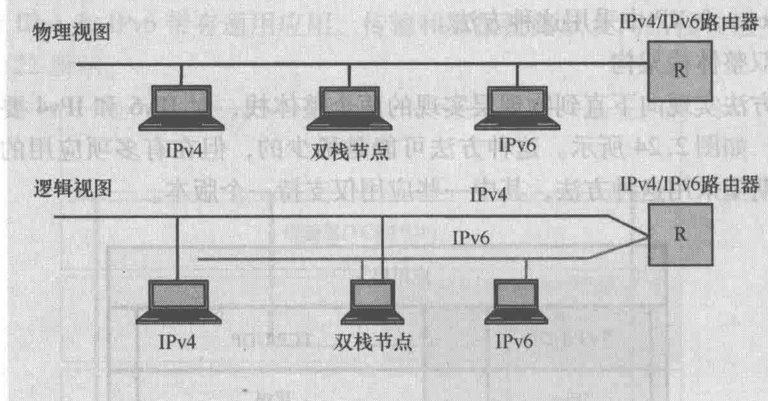


图 2.25 双栈路由器

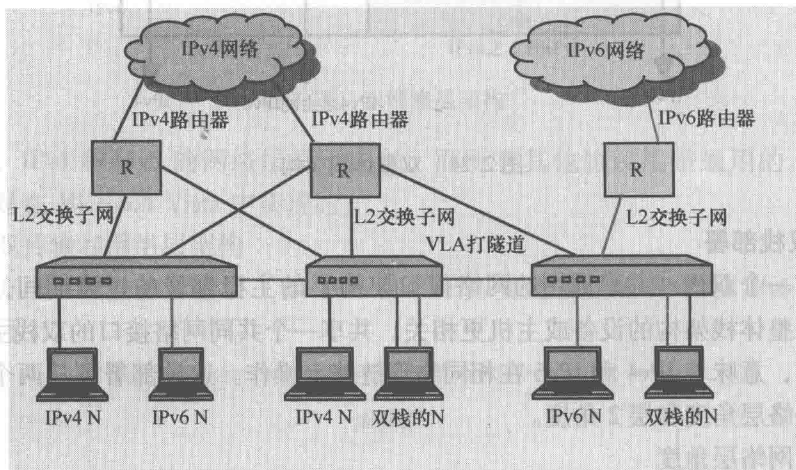


图 2.26 双栈部署 VLAN

2.3.3 打隧道（协议封装）方法

在如下情况下使用这种转换技术，其中完全的基础设施或其各部分，还不能提供原生 IPv6 功能。这项技术的主要优势是它可用在当前 IPv4 基础设施顶部，不必对 IPv4 路由或路由器做出任何重大改变。但是，这要求在选中的路由器或主机处实现双栈，这取决于所用的打隧道方法。

打隧道采用这个过程：来自一个协议的信息或数据被封装在另一个协议的分组内部，由此支持原始数据在使用后一种协议的网络上承载。当使用相同协议的两个节点或网络希望在使用另一种网络协议的一个网络上通信时，可使用这种机制。

打隧道过程涉及三个步骤：封装、解封装和隧道管理。它要求两个隧道端点，

一般而言，是双栈 IP/IPv6 节点（通常是路由器或有时是主机），来处理封装和解封装。

一般而言，通过 IPv4 网络以隧道方式传输 IPv6 分组，要求在隧道的入口点对每条 IPv6 分组封装一个 IPv4 首部，如图 2.27 所示。在这种情形中，IPv4 首部的协议字段被设置为 41（十进制值），指明被封装的 IPv6 分组。同样，这个 IPv4 分组首部的源和目的地址字段分别包含隧道入口点的 IPv4 地址和隧道结束点的 IPv4 地址。这使被封装的 IPv6 分组能够在一个 IPv4 路由基础设施之上被路由。隧道结束点实施解封装，去除 IPv4 首部，并通过 IPv6 网络将 IPv6 分组路由到其目的地[27,32]。

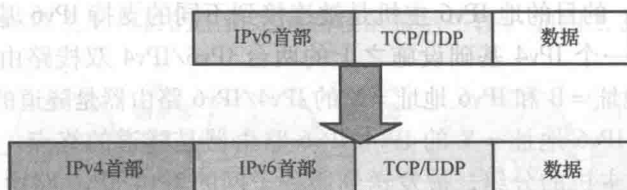


图 2.27 被封装的 IPv6 分组

1. 隧道类型

一般而言，隧道可以有两种类型：配置的和自动的。在配置的打隧道法中，在通信之前，由管理员预定义隧道。例如，基于目的地址和隧道路径参数［如最大传输单元（MTU）］配置隧道端点。另外，自动打隧道法不要求预配置。在这种情形中，在没有人工干预的条件下，在 IPv6 分组中所含信息（如源和目的 IP 地址）的基础上，自动地创建隧道。同样，当讨论从 IPv4 转换到 IPv6 时，也必须考虑两种类型的隧道：IPv4 上的 IPv6 和 IPv6 上的 IPv4。在图 2.20 的阶段 1 到阶段 4 期间，要求 IPv4 上的 IPv6 隧道，而在图 2.20 的阶段 4 期间，则要求 IPv6 上的 IPv4 隧道。将在下面讨论所有这些打隧道技术。注意，因为配置的打隧道法无论如何是以手工方式完成的，在使用范围方面对其形成约束，从技术角度看，自动的打隧道法是非常重要的。所以，从现在开始，我们指出，打隧道法仅意指自动打隧道法。

基于隧道端点定义，RFC 2893（4213）[27,32]定义了 IPv4 上 IPv6 打隧道法配置的四种类型。但是，如前所述，在所有隧道类型中的打隧道过程是保持相同的。注意，开发完成的或正在开发的所有打隧道技术都属于这些类型之一。现在继续讨论这四种在 IPv4 上 IPv6 打隧道法的配置。

（1）路由器到路由器

在路由器到路由器的打隧道配置中，两台 IPv6/IPv4 路由器在一个 IPv4 基础设施上连接两个支持 IPv6 的基础设施。

端点跨越源和目的地之间路径中的一条逻辑链路。在两台路由器之间 IPv4 上的 IPv6 隧道，作为单跳。图 2.28 形象地说明了这个场景。

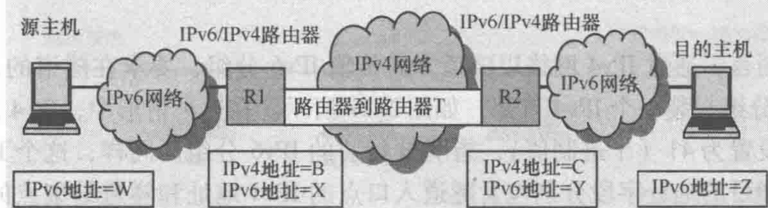


图 2.28 路由器到路由器隧道

在图2.28中，带有IPv6地址=W（在左侧）的源IPv6主机和带有IPv6地址=Z（在右侧）的目的地IPv6主机是被连接到不同的支持IPv6基础设施的。这两个网络是通过一个IPv4基础设施之上的两台IPv6/IPv4双栈路由器而相互连接的。带有IPv4地址=B和IPv6地址=X的IPv4/IPv6路由器是隧道的起点，而带有IPv4地址=C和IPv6地址=Y的IPv4/IPv6路由器是隧道的终点。目的地为带有IPv6地址=Z的主机的分组，被发送到服务子网的IPv4/IPv6路由器（在左侧）。这台路由器配置为以隧道方式传输目的地为如下网络的分组，主机Z驻留在这个网络上，且该路由器以一个IPv4首部封装IPv6分组。这个IPv4首部源字段包含IPv4地址=B，且目的地址包含IPv4/IPv6路由器的IPv4地址=C（在右侧），该路由器被连接到IPv6地址=Z的主机所驻留的IPv6网络。终点IPv4/IPv6路由器（在右侧）解封装IPv4分组，剥离它的IPv4首部，并将原IPv6分组路由到其拟设的目的地。这里是在左侧的IPv4/IPv6路由器起始到右侧IPv4/IPv6路由器的一条隧道，被称作路由器到路由器隧道。

（2）主机到路由器

下面驻留在一个IPv4基础设施内的IPv6/IPv4双栈主机，希望与驻留在IPv6基础设施中的原生IPv6主机通信。为做到这一点，双栈主机创建一条IPv4上的IPv6隧道，到达一台IPv6/IPv4路由器。隧道端点跨越源和目的节点之间路径的第一个分段。IPv6/IPv4双栈源主机和IPv6/IPv4双栈路由器之间的隧道作为单跳。源双栈主机将IPv6分组封装到一个IPv4首部，并将它们发送到双栈路由器，而这台双栈路由器（作为隧道的一个终点）解封装IPv6分组，并以原生方式在IPv6网络中路由这些分组，如图2.29所示。

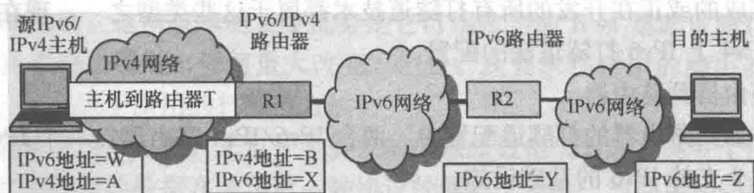


图 2.29 主机到路由器隧道

在图2.29中，给出分组首部地址。除了隧道起点和终点外，隧道机制与路由器到路由器情形是相同的。这里，在左侧隧道的带有IPv6地址=W和IPv4地址=A的源双栈（IPv6/IPv4）主机，（封装）IPv6分组到本地双栈（IPv4/IPv6）路由器（带有IPv4地址=B和IPv6地址=X）。在接收到这条封装的IPv6分组时，隧道终点双栈路由器解封装IPv6分组，并在原生IPv6网络中将之路由到其目的节点（带有IPv6地址=Z）。

（3）路由器到主机

在路由器到主机打隧道配置中，一台（IPv6/IPv4）双栈路由器，在一个IPv4基础设施间创建一条IPv4上IPv6的隧道，到达一台（IPv6/IPv4）双栈主机。这里，隧道端点跨越源主机和目的主机之间路径的最后一段，如图2.30所示。

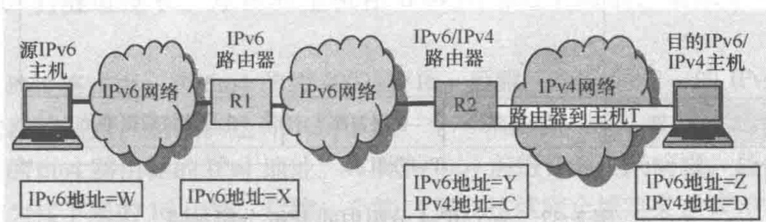


图2.30 路由器到主机隧道

这里，在左侧带有IPv6地址=W的源发IPv6主机，将IPv6分组发送到其本地路由器，该路由器将分组路由到最靠近目的双栈（IPv4/IPv6）主机（带有IPv6地址=Z）的双栈（IPv4/IPv6）路由器。这台服务路由器被配置为在IPv4之上以隧道方式传输IPv6分组到目的地。这里要注意，隧道起点是右侧带有IPv4地址=C的一台双栈（IPv4/IPv6）路由器，它封装IPv6分组，而隧道终点是带有IPv4地址=D和IPv6地址=Z的一个目的主机，由它解封装分组。

（4）主机到主机

在主机到主机打隧道配置中，驻留在一个IPv4基础设施内的一台IPv6/IPv4主机，创建一条IPv4上IPv6隧道，到达另外任何一台IPv6/IPv4主机（也驻留在相同IPv4基础设施内）。隧道端点（起点和终点）跨越源和目的主机之间的整条路径。由此，IPv6/IPv4主机之间的隧道作为单跳。

如图2.31所示，这个隧道配置使源IPv4/IPv6主机通过IPv4网络上的隧道与目的IPv4/IPv6主机通信。这种类型隧道法的其他细节留给读者作为一项练习。当IPv4/IPv6主机希望与一个IPv4基础设施上的应用通过IPv6通信时，使用这种类型的隧道。

2. IPv4上IPv6打隧道法

这种打隧道技术涉及以一个IPv4首部封装IPv6分组，如图2.32所示，从而使IPv6分组可在一个IPv4基础设施之上发送。

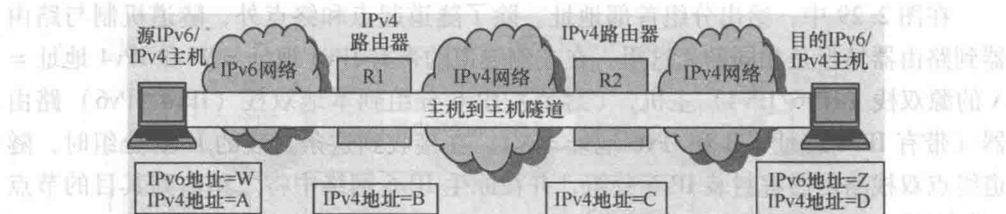


图 2.31 主机到主机隧道

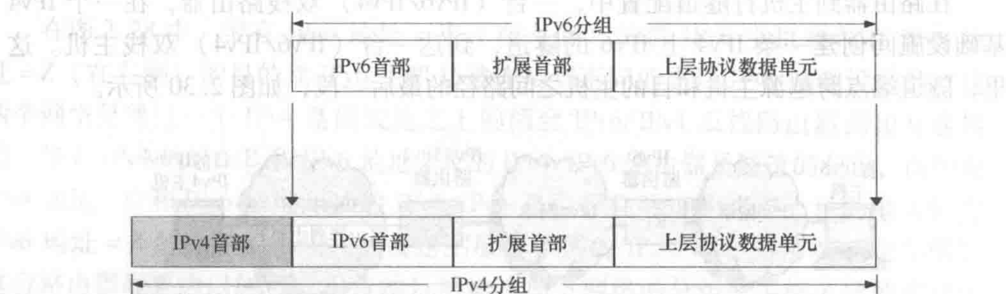


图 2.32 一个 IPv4 分组中的 IPv6 分组封装

这个 IPv4 首部中的如下字段是这样设置的：

1）协议字段：它被设置为41，指明一条被封装的 IPv6 分组。

2）源和目的字段：这些字段分别被设置为起始隧道端点和终止隧道端点的 IPv4 地址。这些是自动地从目的地的匹配路由的下一跳地址和隧道接口推导得到的。

3）分片标志字段：被设置为0，指明一个中间 IPv4 节点不对这条分组实施分片。

现在讨论在参考文献中存在的这种类型的一些打隧道技术。

(1) 6to4

这是基于一个特定全局前缀和内嵌 IPv4 地址的一种自动路由器到路由器打隧道技术（见图 2.28），见 RFC 3056[28] 中的讨论。这项技术依赖于一个特定的 IPv6 地址格式，称作 6to4 地址，来识别 6to4 分组，并据此在一个 IPv4 基础设施上对之打隧道。这个机制使有 6to4 主机（有 IPv4 地址、IPv6 地址和 6to4 地址）的两个站点进行通信，通过连接到共同 IPv4 网络的 6to4 路由器互联。发送主机将使用它的 6to4 地址作为源地址，目的 6to4 地址作为目的地址。在从源 6to4 主机接收到一条分组时，（源）6to4 路由器以 IPv4 首部封装该分组，其 IPv4 地址放在源地址字段，目的 6to4 路由器的 IPv4 地址放在目的字段。在接到分组时，目的 6to4 路由器对之解封装，并在其网络上传输分组，交付到其目的主机，如图 2.33 所示。

注意，这里源和目的主机有 IPv6 地址、IPv4 地址和 6to4 地址（是从其 IPv4 地

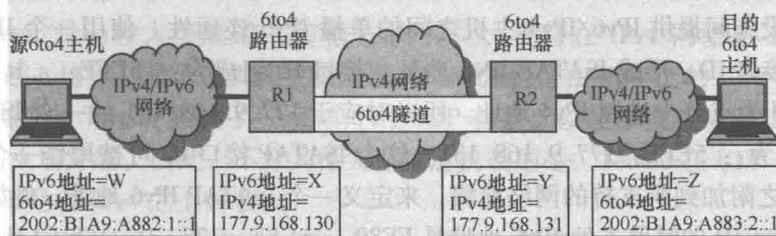


图 2.33　一个 6to4 打隧道的例子

址推导得到的）。对于一个起始隧道端点6to4 路由器和一个终点6to4 路由器，情况相同。6to4 打隧道技术，为 IPv6 主机在 IPv4 网络之上通信，提供了一种高效的机制。

6to4 地址格式由一个6to4 前缀2002::/16，后跟一个全局唯一的 IPv4 地址（用于拟设的目的站点）组成。这种串接形成一个/48 前缀。唯一的 IPv4 地址表示终结6to4 隧道的6to4 路由器的 IPv4 地址。48 比特6to4 前缀作为全局前缀，且一个子网 ID 可附加作为接下来的16 比特，后跟一个接口 ID，这就完全地定义了 IPv6 地址。

例如，考虑一个6to4 网络中具有一个全局唯一 IPv4 地址 177.9.168.130（即站点路由器地址）、子网 ID 1 和接口 ID 1 的一个6to4 网络站点中的一台主机。在6to4 格式中其地址是什么？

6to4 地址格式的前缀将是2002：B1A9：A882::/48，其中 B1A9：A882 是 IPv4 地址 177.9.168.130 的十六进制格式。所以，主机的完整地址将是2002：B1A9：A882：1::1。

（2）6over4

这是一种自动打隧道技术，可以是主机到主机、主机到路由器或路由器到路由器形式，其中相应的主机和路由器被配置为支持6over4。这种方案利用 IPv4 组播，考虑它为虚拟以太网。这意味着主机的 IPv6 地址（6over4 地址格式）是这样形成的，通过使用链路本地范围 FE80::/10 作为网络前缀和主机的 IPv4 地址作为接口 ID。例如，带有 IPv4 地址 193.223.16.6 的一个6over4 将形成其6over4 地址，接口 ID =:: C0DF：1055（以十六进制表示的 IPv4 地址），所以其 6over4 地址为 FE80:: C0DF：1055。

之后，带有6over4 地址的这些 IPv6 分组使用相应的 IPv4 组播地址，被打在 IPv4 首部的隧道内。组播组的所有成员接收打上隧道的分组，且拟设的接收者剥离 IPv4 首部，并处理 IPv6 分组。只要至少一台 IPv6 路由器（也允许6over4）通过 IPv4 组播机制是可达的，则路由器可作为一个隧道端点，并通过 IPv6 路由该分组。

（3）ISATAP

RFC 4214[33] 描述站点间自动隧道寻址协议（ISATAP）作为一种基于地址指派

的主机到主机、主机到路由器和路由器到主机的自动化打隧道技术，该技术在一个IPv4 基础设施间提供 IPv6/IPv4 主机之间的单播 IPv6 连通性。使用一个 IPv4 地址来定义其接口 ID，形成 ISATAP IPv6 地址。接口 ID 组成为：: 5EFE：a. b. c. d，其中 a. b. c. d 为点分十进制 IPv4 地址。所以对应于 177. 9. 168. 131 的一个 ISATAP 接口 ID 表示为：: 5EFE：177. 9. 168. 131。这个 ISATAP 接口 ID 可被用作一个常规接口 ID，将之附加到所支持的网络前缀，来定义一个 ISATAP IPv6 地址。例如，使用 ISATAP 接口 ID 的链路本地 IPv6 地址是 FE80：: 5EFE：177：9：168：131。

（4）隧道代理（broker）

这是 IPv4 网络上的一种自动打隧道技术，其中隧道代理管理来自双栈客户端和隧道代理服务器的隧道请求，隧道代理服务器连接到 IPv6 网络。在一定方面，两个 IPv6 网络之间隧道连接的这个建立过程，看起来类似于建立一条标准的虚拟专用网（VPN）连接。隧道代理可使用 DNS 或证书或这两者实施认证和授权服务。双栈客户端为其隧道端点提供 IPv4 地址、客户端的完全合格的域名（FQDN）、请求的 IPv6 地址数以及客户端是一台主机或一台路由器。一旦授权，隧道代理实施如下任务来代理隧道的创建：

1）在所请求地址数和客户端类型（路由器或主机）的基础上，将一个 IPv6 地址或前缀指派到客户端。

2）在 DNS 中注册客户端 FQDN。

3）指派和配置一个隧道服务器，并将客户端的被指派的隧道服务器及关联隧道和 IPv6 参数（包括地址前缀和 DNS 名）通知客户端。

图 2. 34 形象地说明一个双栈客户端、隧道代理交互以及客户端和被指派隧道服务器之间一条隧道的建立。

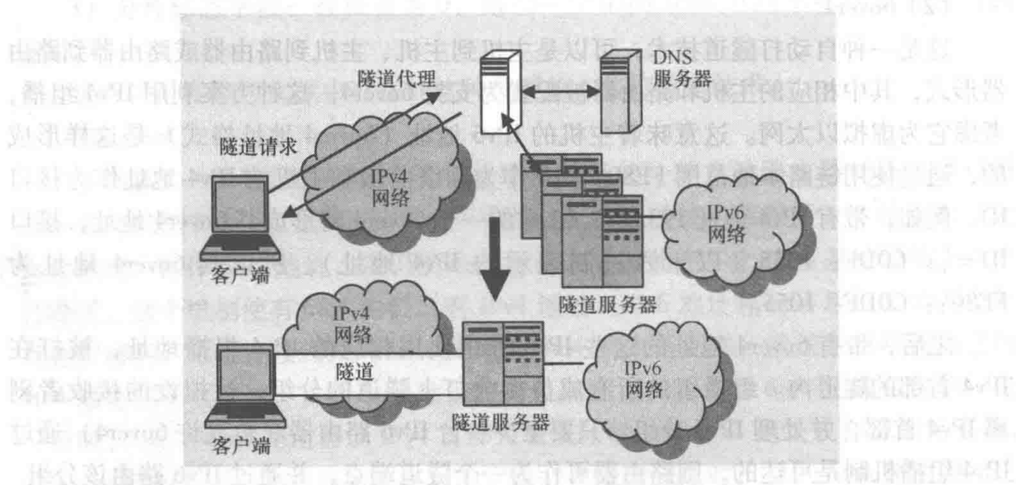


图 2. 34 隧道代理系统

(5) Teredo

Teredo 是一种大型隧道(传输)技术[36],支持在 IPv4 网络上 IPv6 分组的 NAT 穿越。这里,对于主机到主机隧道,IPv6 分组是在 IPv4 上的 UDP 之上以隧道方式传输的。Teredo 集成了一个额外的 UDP 首部,以便于支持 UDP 端口转换的 NAT/防火墙穿越。由此,不像其他大型隧道机制的是,Teredo 在 UDP 中封装 IPv6 分组而不是在 IPv4 之上封装。

如在 RFC 4380 中定义的一样,Teredo 要求如下元素,如图 2.35 所示。

- Teredo 客户端。
- Teredo 服务器。
- Teredo 中继(TR)。

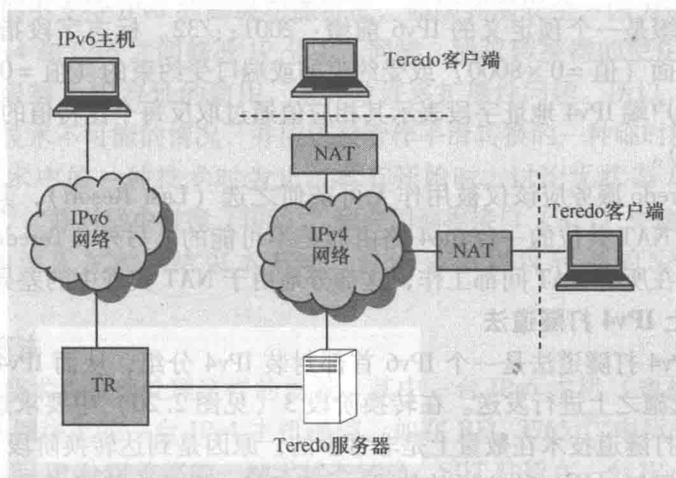


图 2.35 Teredo 系统

Teredo 主机/客户端预配置有要使用的 Teredo 服务器 IPv4 地址。Teredo 打隧道过程由两项任务组成:首先识别最接近拟设目的 IPv6 主机的 TR 和之后识别起作用的 NAT 防火墙类型。为了针对一个给定目的 IPv6 主机确定最近的 TR,Teredo 客户端向目的主机发送一条 ping [因特网控制消息协议版本 6(ICMPv6)回声请求]。ping 被封装 UDP 首部和 IPv4 首部,并被发送到 Teredo 服务器。一个著名的 UDP 端口 3544 由 Teredo 服务器使用,侦听来自 Teredo 客户端的请求。Teredo 服务器解封装并通过 TR 发送原生 ICMPv6(因特网控制消息协议版本 6)分组到拟设的目的主机。接下来,目的主机的响应将通过原生 IPv6 到 TR,之后到 Teredo 服务器,被路由回到原主机。在这种方式中,客户端确定合适 TR 的 IPv4 地址和端口。

一般而言,NAT 设备将来自同一内部 IP 地址和端口的所有分组映射到一个对应的外部地址和端口,从而内部主机可与外部主机通信,反之亦然,这被称作全锥面(Full-Cone)NAT。在受限锥面 NAT 和端口受限 NAT 设备的情形中,如果内部

主机已经向外部主机发送过一条分组，或分别采用相同的主机地址和端口号向外部主机发送过一条分组，则外部主机仅可与内部主机通信。所以，为在这两种情形中的双向通信完成 NAT 映射，则 Teredo 客户端发送一个气泡（Bubble）分组到外部主机，该分组是没有净荷的一个 IPv6 首部。

Teredo IPv6 地址[35,36]有如图 2.36 所示的格式。

比特	32	32	16	16	32
字段	Teredo 前缀	Teredo服务器 IPv4地址	标志	客户端端口	客户端 IPv4地址

图 2.36 Teredo 地址格式

Teredo 前缀是一个预定义的 IPv6 前缀：2001::/32。标志字段指明 NAT 的类型，因为全锥面（值 = 0 × 8000）或受约束的或端口受约束的（值 = 0 × 0000）。客户端端口和客户端 IPv4 地址字段表示其相应值通过取反每个比特值的混杂值（Obfuscated Value）。

如此，Teredo 服务应该仅被用作无可奈何之选（Last Resort），其中直接 IPv6 连接能力或与 NAT 共位的一台 6to4 路由器是不可能的。另外，Teredo 方法是复杂的，不能保障在所有 NAT 间都工作，这部分地由于 NAT 实现中的差异导致的。

3. IPv6 上 IPv4 打隧道法

IPv6 上 IPv4 打隧道法是一个 IPv6 首部封装 IPv4 分组，从而 IPv4 分组可在一个 IPv6 基础设施之上进行发送。在转换阶段 3（见图 2.20）中要求这种类型的打隧道法。这种打隧道技术在数量上是非常少的，原因是到达转换阶段 3，还有很长的路要走。这里仅讨论一种这样的技术——DSTM，只是为使读者了解这种技术的存在。

双栈转换机制（DSTM）提供了以隧道方式在 IPv6 网络上传输 IPv4 分组（最后到达目的 IPv4 网络和主机）的一种方式。在 IPv6 网络上意图与 IPv4 主机通信的主机，将要求一个 DSTM 客户端。在将拟设目的主机的主机名仅解析到 IPv4 地址时，客户端将发起 DSTM 过程，这类似于隧道代理方法。

该过程开始于 DSTM 客户端联系一台 DSTM 服务器，以便得到一个 IPv4 地址（首选通过 DHCPv4）和 DSTM 网关的 IPv6 地址。IPv4 地址被用作要被传输的数据分组中的源地址。这个分组被封装一个 IPv6 首部，使用 DSTM 客户端的源 IP 地址和 DSTM 网关的 IPv6 地址作为目的地。注意，这里，DSTM 客户端是 IPv6 上 IPv4 隧道的起点，而 DSTM 网关是隧道的终点。IPv6 首部中的下一字段指明一条被封装的 IPv4 分组。

2.3.4 转换方法

为使 IPv4 向 IPv6 平滑地转换，转换技术形成一个至关重要的部分（Lot），虽

然相比其他技术，就扩展性和安全性而言，它有一些劣势。在协议栈的一个特定层，典型情况下是网络、传输或应用层，转换技术实施 IPv4 到 IPv6 转换，并进行相反转换。某时可能要求这些技术的场景如下：

1）一台 IPv6 主机希望与一个 IPv4 基础设施之上的一台 IPv4 主机通信，相反情况依然，这典型地是在转换的第一阶段和第三阶段期间的情况（见图 2. 20）。

2）一台纯 IPv6 主机希望与一台纯 IPv4 节点通信，这典型地是在转换的第二阶段和第三阶段（见图 2. 20）。

3）IPv4 应用需要用于 IPv6 网络上，典型地是在转换的第三阶段和最后一个阶段（见图 2. 20）。

在本节提出讨论这些转换技术中的一些技术，涵盖上述场景。转换方法的主要劣势是，它们不支持 IPv6 的高级功能特征，如端到端安全，因为这些机制确实要双向地在 IPv4 和 IPv6 之间修改 IP 分组。同样，协议转换器的使用导致 NAT 的问题，也高度限制了 IP 寻址的使用，这导致许多扩展性问题。所以，转换技术应该仅用于其他技术不可能的情况，并应该被看作平滑转换的一种临时解决方案，直到实现其他技术中的一种技术时为止。本节开始时，讨论无状态 IP/ICMP 转换（SIIT）算法[24]，并对栈中碰撞（BIS）[26]和应用编程接口（API）中碰撞（BIA）[29]技术抛出一些深入见解。结束本节时，讨论采用协议转换的 NAT（NAT-PT）技术。

1. SIIT 算法

这项转换技术典型地处理这样的场景，其中一台 IPv6 主机（也指派一个 IPv4 地址）与 IPv4 网络上的一台 IPv4 主机通信。如在 RFC 2765[24]中所述，SIIT 提供 IPv6 和 IPv4 之间 IP 分组首部的一种无状态转换。SIIT 驻留在一台 IPv6 主机（它也被指派一个 IPv4 地址），并将外发的 IPv6 分组首部转换为 IPv4 首部，将到达的 IPv4 首部转换为 IPv6 首部。注意，SIIT 规范详细描述了协议转换机制，但没有讨论 IPv4 地址指派到 IPv6 主机的规程。所以，典型情况下，SIIT 被包装在后来开发的像 BIS 和 BIA 的转换技术中，它们提供 IPv4 地址指派规程。现在继续解释 SIIT 算法是如何准确地工作的（见图 2. 37）。

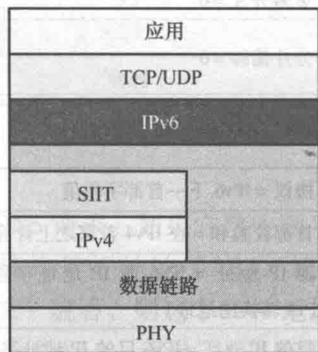


图 2. 37　TCP（UDP）/IP 栈中的 SIIT

假定一台 IPv6 主机在一个 IPv4 网络上与一台 IPv4 主机通信。驻留在 IPv6 主机上的 SIIT 算法，将外发 IPv6 分组首部转换为 IPv4 首部格式，将到达的 IPv4 分组首部转换为 IPv6 分组首部。当 IPv6 首部中的目的 IP 地址是一个 IPv4 映射的地址[35]时，SIIT 算法识别出这样一种情形。转换被解析 IP 地址（如何指派到 IPv6 主机在 RFC 2765 中是没有规定的）到一个 IPv4 映射地址（在图 2. 38 中给出格式）机

制是由后来开发的像BIS和BIA的转换机制提供的。

比特	80	16	32
字段	0	FFFF	IPv4地址

图2.38 IPv4映射的地址格式

在IPv4映射地址作为目的IP地址存在的基础上，SIIT算法实施首部转换，通过数据链路和物理层为转换得到IPv4网络上的一条IPv4分组。欲了解首部转换的细节，请参见表2.4。注意，IPv6首部（被转换为IPv4首部）中的源地址使用与一个IPv4转换格式[35]（同样，一个IPv4地址到一个IPv4转换地址的变换法，超出了RFC 2765的范围）不同的一种格式（见图2.39）。

比特	64	16	16	32
字段	0	FFFF	0	IPv4地址

图2.39 IPv4转换的格式

表2.4 IPv4和IPv6首部转换过程

IPv6到IPv4首部转换	IPv4到IPv6首部转换
版本=4	版本=6
首部长度=5（没有IPv4选项）	流量类=IPv4首部TOS比特
服务类型=IPv6首部流量类字段	流标签=0
净荷长度=IPv4首部总长度值-（IPv4首部长度+IPv4选项长度）	净荷长度=IPv4首部总长度值-（IPv4首部长度+IPv4选项长度）
总长度=IPv6首部净荷长度字段+IPv4首部长度	标识=0
标识=0	下一首部=IPv4首部协议字段
标志=不分段=1，更多分片=0	跳限制=IPv4 TTL字段值-1
分片偏移=0	源IP地址=0:0:0:0:FFFF::/80与IPv4首部源IP地址串接
TTL=IPv6跳限制字段值-1	目的地IP地址=0:0:0:0:0:FFFF::/96与IPv4首部目的地串接
协议=IPv6下一首部字段值	
首部校验和=在IPv4首部之上计算	
源IP地址=IPv6源IP地址字段的低32比特（IPv4转换的地址）	
目的IP地址=IPv6目的IP地址字段的低32比特（IPv4映射的地址）	
选项=无	

这是因为依据 RFC 4213[32]，作为隧道法的源地址，IPv4 映射的地址格式是无效的。因此，它用作源地址，将使通过任何所涉及（intervening）隧道的通信变得无效。使用 IPv4 转换的格式，就旁路了这种可能。所以，一个 IPv4 转换地址指代源 IPv6 节点，而一个 IPv4 映射地址被用来指代目的 IPv4 节点。这样一个 IPv6 首部由 SIIT 转换为一个 IPv4 首部。

另外，从数据链路层到达 SIIT 的分组将其首部从一个 IPv4 转换为 IPv6 首部和地址被转换回 IPv6（IPv4 转换的和 IPv4 映射的）地址。在表 2.4 中汇总了两个方向（即 IPv6 到 IPv4 和 IPv4 到 IPv6）的基本首部转换。

因为 SIIT 是无状态的，所以在一台 IPv6 主机和一个 IPv4 网络之间可能存在许多转换期，一台 IPv6 主机和一台 IPv4 主机之间的分组可通过任意数量的转换器。不需要将每个会话绑定到一个特定的转换器。SIIT 涵盖 IPv6 和 IPv4 之间尽可能多的转换，但不涵盖任何 IPv4 选项和一些 IPv6 选项扩展。SIIT 涵盖从 IPv6 到 IPv4 的 ICMP 消息，也涵盖相反方向的 ICMP 消息。

2. 栈中隆块

对于初始转换阶段，RFC 1993[23] 规范了转换机制，如双栈和隧道法，见前面各节中的细节讨论。这得到支持这些技术的主机路由器的开发。但相比 IPv4，IPv6 存在非常少的应用。为平滑地推进转换，人们期望与 IPv4 相伴（at par with），增加 IPv6 应用的可用性。不幸的是，预计这会用掉一个非常长的时间。在这方面，RFC 2767[26] 为双栈主机提出称作 BIS 的一种机制，这使使用 IPv4 应用的双栈主机在 IPv6 网络上进行通信。这项技术插入窥探（Snoop）TCP/IPv4 层和链路层（如网卡）之间数据流的模块，并将 IPv4 分组转换为 IPv6 分组，反之亦然。当这些主机在 IPv6 网络上通信时，将池式 IPv4 地址内部指派到这些主机并使用 DNS 协议，这些 IPv4 地址从来就不会流出主机。RFC 2767[26] 规范实施这种转换的三个模块：扩展名解析器（ENR）、地址映射器（AM）和转换器（见图 2.40）。

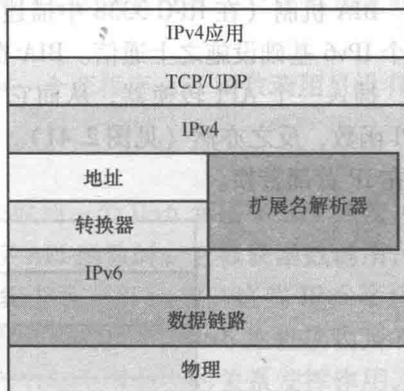


图 2.40 BIS 组件

（1）扩展名解析器

这被实现为 DNS 栈的组成部分。它服务应用的 DNS 请求。如果它接收到一个 IPv4 地址（A 记录）的一条 DNS 请求，则它为所附接的 IPv6 地址（AAAA 记录）产生一条 DNS 请求。如果 DNS 服务器以一条"A 记录"应答，则这个 IPv4 地址仅被转发到 IPv4 连接的应用。如果 DNS 响应仅包括一条"AAAA 记录"，则 ENR 导致 AM 将这个 IPv6 地址附接到一个临时的 IPv4 地址。之后，它将这个"A 记录"发回应用。注意，这里因为这些地址仅可内部使用，所以可能使用一个私有 IPv4

地址空间。

(2) 地址映射器

AM维护一个IPv4 地址池（Spool）。该池由一个隐私地址空间组成。同样它在一个表中存储一个临时IPv4 地址和IPv6 地址之间的关系。它分别在如下情形中由ENR 或转换器使用：

1）如果ENR 仅接收到一条AAAA记录且不存在关系。

2）如果转换器接收到没有找到关系的一条IPv6 分组。

注意，当初始化该表时，仅存在一种意外情况：它静态地将一对IPv4 地址和IPv6 地址注册到该表。

(3) 转换器

使用SIIT 机制中规范的算法[3]，这个模块将IPv4 分组转换为IPv6 分组，相反情况亦然。当从IPv4 应用接收到IPv4 分组时，它将IPv4 首部转换为IPv6 首部，并分片IPv5 分组，原因是IPv6 首部长度典型地比IPv4 首部长度要多20 字节，从而使分组尺寸不能超过IPv4 网络的分组最大传输单元（PMTU），并将分片发送给IPv6 栈。当从IPv6 网络接收到IPv6 分组时，这个模块以对称于前面的情形而发挥作用，它将IPv6 分组转换回IPv4 分组，例外情况是，不需要对分组进行分片。

3. API 中碰撞

BIA 机制（在RFC 3338 中描述）[29]，像BIS，支持IPv4 应用的使用，同时在一个IPv6 基础设施之上通信。BIA 在双栈主机的套接字API 模块和TCP/IP 模块之间，插入一个API 转换器，从而它可将IPv4 套接字API 函数转换为IPv6 套接字API 函数，反之亦然（见图2.41）。注意采用这种机制，该转换得以简化，因为不存在IP 首部转换。

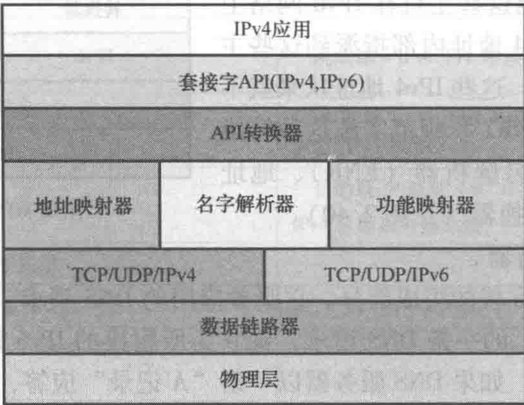


图2.41　BIA 组件

当双栈 IPv6 主机上的 IPv4 应用与其他 IPv6 主机通信时，API 转换器检测到来自 IPv4 应用的套接字 API 函数，并触发 IPv6 套接字 API 函数与 IPv6 通信，反之亦然。由此，双栈主机（参考文献［29］）有一个 API 转换器，使用一个 IPv6 基础设施上现有 IPv4 应用与其他 IPv6 主机通信。API 转换器由三个模块组成：名字解析器、地址映射器和函数映射器。

（1）名字解析器

作为对一条 IPv4 应用请求的响应，名字服务器返回一个合适的应答。当一个 IPv4 应用发送针对一条“A 记录”的一条 DNS 请求时，名字解析器截获该请求，并创建一条新的查询，请求“A”和“AAAA”记录。如果 DNS 以一条“A 记录”做出应答，则这个 IPv4 地址仅转发到一条 IPv4 连接的应用。如果 DNS 应答仅包括一条“AAAA 记录”，则名字解析器使 AM 将这个 IPv6 地址附接到一个临时 IPv4 地址。之后名字解析器将这条“A 记录”发回应用。

（2）地址映射器

地址映射器（AM）内部维护一个 IPv4 地址和一个 IPv6 地址对的一个表。IPv4 地址是从一个 IPv4 地址池中指派的。它使用 0.0.0.1 和 0.0.0.225 之间的未指派地址。当名字解析器或函数映射器请求 AM，指派对应于一个 IPv6 地址的一个 IPv4 地址时，它从池中选择并返回一个 IPv4 地址，并在表中动态地注册一个新表项。注册发生在如下两种情形中：

1）当名字解析器仅得到目标主机名字的 AAAA 记录且没有一个 IPv6 地址的映射表项时。

2）当函数映射器从所接收到的数据中得到一个套接字 API 函数调用且没有 IP 源地址的映射表项时。

（3）函数映射器

函数映射器将一个 IPv4 套接字 API 函数映射到一个 IPv6 套接字 API 函数，反之亦然。在检测到来自 IPv4 应用的 IPv4 套接字 API 函数时，它截获函数调用，并触发对应于 IPv4 套接字 API 函数的那些 IPv6 套接字 API 函数。这些 IPv6 套接字 API 函数将被用来与目标 IPv6 主机通信。类似地，在从其他 IPv6 主机接收到的数据中检测到 IPv6 套接字 API 函数时，它以对称于前一情形中的关系发挥作用，并将它们转换回 IPv4 套接字 API 函数。

4. 网络地址转换-协议转换

定义于 RFC 2766[25,38] 中的 NAT-PT，本质上是在纯 IPv6 和纯 IPv4 节点之间通信的一种方法。它支持原生 IPv6 主机和应用与 IPv4 主机和应用通信，反之亦然。存在驻留在 IPv6 和 IPv4 网络之间边界处的一台 NAT-PT 设备，有利于这些 IP 异构主机之间的通信。一台 NAT-PT 设备有利于纯 IPv6 和纯 IPv4 主机之间的通信，方法是维护每条连接的状态，并为每个会话实施地址转换和协议转换。不需要客户端配置，且所有 NAT-PT 转换都是在 NAT-PT 设备处完成的，对端用户是完全透明

的。所以，如此说，网络层安全是不能保障的。同样，由于IPv4和IPv6首部格式之间的显著差异，这种方法仅能以一种尽力而为的方法完成。NAT-PT将一条IPv4数据报转换到一个语义上等价的IPv6数据报，或相反情况（见图2.42）。NAT-PT的NAT部分将一个全局可路由IPv4地址转换为一个全局可路由IPv6地址，或相反情况。NAT-PT的PT部分处理语义上等价的IP首部的解释和转换，从IPv4到IPv6或相反情况，见SIIT机制中的描述。像NAT一样，NAT-PT也使用一个地址池，它动态地指派到被转换的数据报。

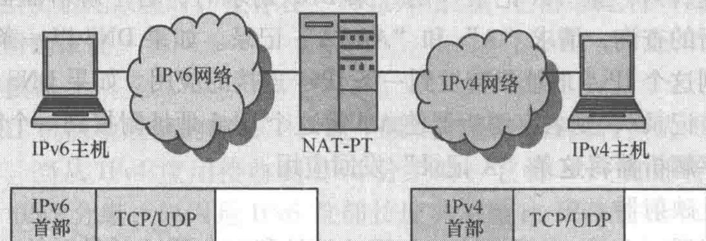


图2.42 NAT-PT系统

NAT-PT可扩展到网络地址和端口转换-协议转换（NAPT-PT）。NAPT-PT将地址转换推进一步，方法是也支持端口号的转换。这使之重用一个IPv4池地址并映射到多台主机成为可能。NAT-PT不能处理在一条IP分组内带有内嵌IP地址的IP应用，原因是它不能查看净荷内部以便转换那些IP地址。RFC 2766在一台NAT-PT设备内规范了称作应用层网关（ALG）的另外实体，来处理这些应用。一个ALG查看净荷内部，并转换那些IP地址。为支持诸如DNS和FTP等应用，ALG是必要的。

2.4 路由

在因特网中，为支持没有直接连接的各主机进行通信，路由是一项至关重要的功能。路由的主要功能是在所连接网络分段（链路或子网）之间转发分组的过程。使这种情况发生的网元是路由器。在IPv4中，对于分组路由，层次结构的第一层表示为各子网，其中每台主机是直接连接的。在发送分组之前，每台源主机都要做一项测试，确定目的主机是在线的（在与源相同的子网中）或离线的（不在相同子网中）。在第一种情形中，源直接将分组发送到目的地；在第二种情形中，源将分组转发到子网上的路由器，该路由器通过咨询其路由表，确定哪个是去往一个给定目的地的最佳路径，并将该分组转发到下一台合适的路由器。在路由算法和路由协议（正用于传输之中）的基础上，形成这些路由表。在源和目的主机是如何放置在因特网的基础上，确定这些路由协议。如此，为高效的路由，在IPv4中开发

了许多路由协议。在 IPv6 中的路由几乎等同于无类域间路由（CIDR）下的 IPv4 路由，例外是在 IPv6 中地址是 128 比特的，而 IPv4 地址是 32 比特的。采用非常直接的扩展，所有 IPv4 路由算法，如开放最短路径优先（OSPF）、路由信息协议（RIP）、域间路由协议（IDRP）和中间系统到中间系统（IS-IS），可被用于 IPv6 中的路由。IPv6 中的这些简单扩展，也包括对新的强大路由能力的支持，如基于策略、性能和成本的提供商选择；主机移动性；多穴连接路由域；自动重新寻址；路由和当前位置。在本节，以这样的方式开始讨论，首先讨论 IPv6 网络架构，之后继续讨论理解路由核心部分，这是各种路由协议的基础。在此之后，将讨论焦点放在一些主要路由协议上，这是从其对 IPv6 扩展的角度讨论的。

2.4.1　网络架构

在前面有关寻址小节指出，所有类型的 IPv6 地址都被指派到接口，而不是节点或主机。如此，每个节点属于单个节点；节点的或主机的任何接口的单播地址可被用作该节点的一个标识符。考虑到这个事实，通用 IPv6 网络架构如图 2.43 所示。

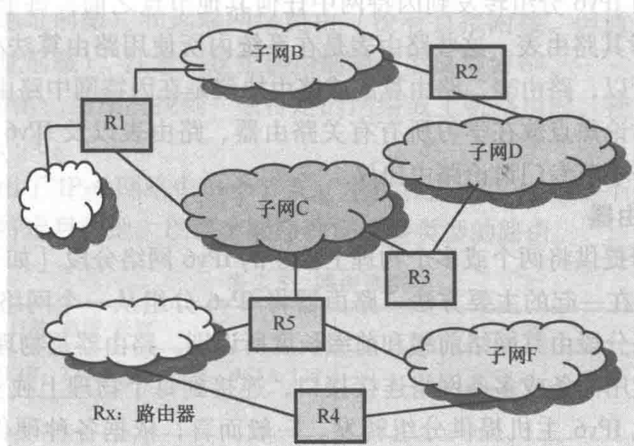


图 2.43　IPv6 网络架构

这个去中心化的网络架构由许多通过路由器互联的子网组成。接下来，这些子网被归组为受控子网的集合，并由称作自治系统（AS）的一个唯一权威所管理。在同一 AS 内路由消息的路由器被称作内部路由器，而那些在不同 AS 之间路由分组的路由器被称作外部路由器。在两个 AS（由 A 和 B 指明）之间互联的一个例子如图 2.44 所示。

内部路由器通过一种内部网关路由协议（IGRP）交换路由信息，而外部路由器使用一种外部网关路由协议（EGRP）交换路由信息。正常情况下在一个 AS 内的所有路由器中使用相同的 IGRP。

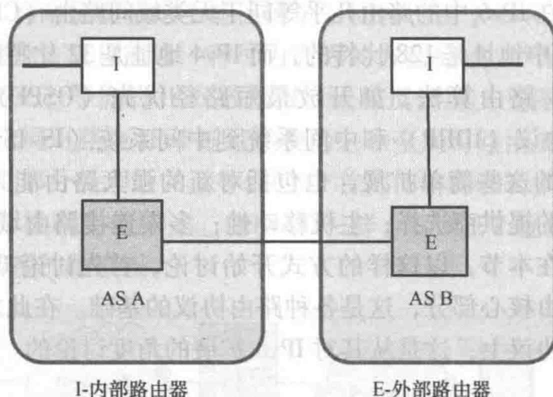


图2.44 两个AS之间的互联

2.4.2 路由核心知识

路由器是这样的网元，它们主要负责在因特网间（源和目的主机之间）路由分组。在将任何IPv6分组转发到因特网中任何其他节点之前，每个节点（主机或路由器）要参考其路由表。这些路由表是在系统内所使用路由算法/协议的基础上进行维护的。所以，路由表、路由算法或路由协议是在因特网中路由一条分组的基础。在本节将讨论焦点放在学习所有有关路由器、路由表以及IPv6中使用的路由算法上，而下一小节专门讨论路由协议。

1. IPv6路由器

IPv6路由器提供将两个或多个物理上独立的IPv6网络分段［如子网和AS（见图2.43）］连接在一起的主要方法。路由器将IPv6分组从一个网络分段转发到另一个分段。网络分段由其网络前缀和前缀长度所识别。路由器是物理上多穴连接的主机，即它们使用两条或多条网络连接接口，连接到每个物理上独立的网络分段，这些分段为其他IPv6主机提供分组转发。一般而言，依据各种硬件和软件系统，由其建立情况区分各路由器。高速路由器是运行特定软件的专用硬件设备，在IPv6网络架构中是比较常见的。不管其硬件和软件配置为何，所有路由器在它们转发其他通信主机分组的基础上，维护路由表。

2. 路由表

IPv6节点（主机和路由器）使用路由表，维护有关其他IPv6网络和节点的信息。一个路由表提供有关与远端网络分段和主机方面的有用信息。运行IPv6的每个节点或设备，在IPv6路由表的基础上，确定如何转发分组。它特别地存储有关IPv6地址前缀的信息以及它们是如何可达的，直接的还是间接的。

当IPv6初始化时，IPv6路由表项是默认创建的，在接收到包含在线前缀和路由表的RA消息时，添加额外表项。在检查IPv6路由表之前，为匹配IPv6分组

（正被转发的）中目的地址的一个表项检查目的地缓存。如果在目的地缓存中没有找到表项，那么使用路由表确定下一跳接口（为转发分组要使用的物理接口）和下一跳地址（路由器的地址）。并据此更新目的地缓存，从而要被转发的后续分组使用目的地缓存表项。下面是一个典型 IPv6 路由表项的各字段：

1）目的地前缀：目的地前缀是一个 IPv6 地址前缀，可有从 0 到 128 的一个前缀长度。

2）下一跳地址：这是分组要被转发到的地址。

3）接口：使用该网络接口转发该分组。

4）度量：这是一个数字，它被用来指定路由的成本，从而可为一个特定目的地在多条路由间选择最佳路由。

典型情况下，可使用路由表项存储如下路由类型的信息：

1）主机特定路由：这是到一个特定 IPv6 地址的一条路由。主机路由支持在每 IPv6 地址基础上进行路由。注意，对于主机路由，路由前缀是带有前缀长度为 128 的一个特定 IPv6 地址。

2）网络特定路由：在这种情形中，有两个类型，即直接附接的网络路由（直接附接子网的地址前缀）和远端网络路由（没有直接附接，但通过其他路由器是可达子网的地址前缀）。注意，这里路由前缀长度典型为 64。

3）默认路由：当没有找到一条特定的网络或主机路由时，使用默认路由。默认路由前缀为::/0。

表 2.5 给出了 IPv6 网络中一个节点（主机和路由器）的一个典型路由表。注意，对于一个特定目的地，以那个顺序确定这些类型的路由。

表 2.5　路由表例子

目的/前缀-长度	下　一　跳	接　口
2001：DB8：0：2F3B：2AA：FF：FE28：9C5A/128（主机特定路由）	路由器 1	A
2001：DB8：0：2F3B：：/64（网络特定路由）	路由器 2/“直接的”	B
：：/0（默认路由）	路由器 3	A

路由确定过程

IPv6 遵循如下步骤，确定转发一条分组要使用哪条路由表项：

1）对于发送节点，如果源地址是由发送应用指定的，那么仅检查指派到源地址的接口 ID 的那些路由。

2）对于发送主机，如果源地址没有由发送应用指定，那么检查所有路由。

3）对于每条被检查的路由表项，对于在路由前缀长度中指定的比特数，IPv6 将网络前缀中的比特与目的地址中的那些比特进行比较。

4）对于一个表项，如果网络前缀中的所有比特匹配目的地址中的所有比特，

那么该路由就是目的地的一个匹配。

5）收集匹配路由的一个列表。选择具有最大前缀长度的路由（匹配目的地址最高位比特的路由）。最长匹配路由是到目的地的最具体的路由。如果找到具有最长匹配的多个表项，则路由器使用最低度量选择最佳路由。

路由确定过程的结果是在路由表中选择单条路由。被选中的路由得到下一跳接口和匹配路由。对于远端流量，下一跳地址是存储在下一跳地址字段中的地址（它典型地是一个邻接路由器的地址）。对于去往一条直接附接链路上的各邻居的流量，下一跳地址是分组的目的地址。在这种情形中，一个地址没有存储在下一跳地址字段中。如果在一个发送主机中的路由确定过程不能定位一条匹配路由，那么IPv6将目的地看作是本地可达的。

3. 路由算法

路由协议是在路由算法基础上以合适度量开发得到的。一个路由算法指这样一种方法，协议用其确定任何网络对之间的最佳路由，并在路由器之间共享路由信息。一个度量是成本的一个度量方法，被用来评估一条特定路由的效率。在当前路由协议中最普遍使用的有两种路由算法，即距离矢量和链路状态。存在一些协议，它们使用这些算法与其他类型算法的一个组合体。

（1）距离矢量（Bellman-Ford）路由算法

距离矢量路由算法确定网络上任何两个节点之间的路由，是在其双向距离作为一个度量的基础上确定的。距离度量是跳数，即沿从源节点到目的节点的路径上路由器的数量。除了路由表之外，在网络上的每台路由器维护称作距离向量的一个数据结构。距离矢量为每个目的地包含一个表项，且每个表项包含一个目的地址和关联的跳距离度量。每台路由器周期性地将其距离矢量发送到其邻居路由器，并计算其路由表，合并其活跃邻居的所有距离向量。在从其邻居路由器中接收到距离矢量时，它更新自己的距离矢量，并重新计算它的路由表。这个合并过程是基于最小度量准则的。对于每个目的地，所选中的路径是在所有可能路径中带有最小度量的那条路径。

这种类型的路由表是容易实现的，但却是高复杂性的，在最坏情形中是指数的，正常处在$O(n^2)$到$O(n^3)$的范围，其中n是网络中的节点数量。这使这种算法的使用不适用于1000个以上节点的路由。这种算法的另一个问题是缓慢收敛到稳态路由。该算法以正比于网络上最慢路由器的速度收敛。这个算法被用来计算RIP和IGRP中的路由表。这个算法的实际工作过程及其复杂性分析超出了本章的范围。

（2）路径矢量路由算法

路径矢量算法类似于距离矢量算法，但它不通告度量，相反它通告到达每个目的地要穿越的AS列表。一个AS列表的使用帮助发现网络中的可能环路，由此简化倾向某些路由的路由策略的实现。路径矢量算法用在EGRP之中。

（3）链路状态路由算法

在一个链路状态算法中，每台路由器维护一个映射（Map），通过与网络上其他路由器交互，描述网络的当前拓扑。使用这个映射，通过使用 Dijkstra 算法，路由器计算最佳路由。典型情况下，通过交换链路状态分组（LSP），每台路由器与网络上的其他路由器通信，LSP 提供路由器当前所连接子网的链路状态。每台路由器也维护称作 LSP 数据库的一个数据库，其中存储网络上由其他路由器最近产生的 LSP。LSP 数据库是一个网络图的一种表示，存储为一个邻接矩阵。注意，依据定义，LSP 数据库准确地等同于网络上的所有路由器。由此，带有相关联度量的 LSP 数据库，为一台路由器计算一个路由表提供必要的和充足的信息。链路状态算法的计算复杂度是 O（$L\log N$），其中 L 是网络中的链路数；N 是网络上的节点数。这种类型的算法可部署到大型网络中，且它们动态地适应变化的互联网络状况。同样，这些算法也允许在更真实成本度量的基础上选择路由，而不是网络之间简单的跳数。但是，要建立这样的算法是比较复杂的，且相比距离矢量算法，使用更多计算机处理资源，而且是不太容易稳定的。同样，这些算法的算法细节和复杂度分析超出了本章的范围。

注意，距离矢量路由器仅将有关所有子网的信息发送到它们的邻居路由器，而链路状态路由器将有关子网的信息发送到网络上它们直接连接到的所有路由器。一些链路状态路由协议有 OSPF、IS-IS 和增强的内部网关路由协议（EIGRP）。

2.4.3 路由协议

路由协议使路由器可动态地通告并学习路由，并确定哪些路由是可用的，以及哪些是到一个目的地的最高效路由。路由协议也提供层 3 网络状态更新，并在路由器中在层 3 上传播路由。但是，IP 作为一个层 3 协议，不仅传播路由表，而且通过其分组在网络间传输数据。IPv6 中的路由几乎类似于 CIDR 中的 IPv4，例外是，地址是 128 比特 IPv6 地址，而不是 32 比特 IPv4 地址，虽然在一些协议中，添加了 IPv6 的特定功能，使它们更加鲁棒可靠。对动态 IPv4 路由协议［OSPF、IDRP、RIP、IS-IS、边界网关协议（BGP）］做出最小修改，以可操作 IPv6 地址格式。IPv6 有带有一个改进源路由选项的路由首部（RH），主要引入来包括提供商选择和移动性。这使数据报的发送者，指定在到目的地的路上要访问的地址列表，这非常类似于 IPv4 选项松散/严格源路由，但没有它的重要限制，像首部尺寸和低效率。注意，在 IPv6 RH 中，数据报目的地首部由列表中的下一个地址所替换。一般而言，在 IPv6 中的路由协议被粗分为两类，即静态的和动态的。IPv6 有两种类型的动态路由协议，即内部网关协议（IGP）和外部网关协议（EGP）。动态路由协议 RIPng、EIGRP、因特网协议版本 6 的 OSPF（OSPFv3）和 IS-ISv6 属于 IGP，而动态路由协议 MP-BIG-4 属于 EGP。本节将讨论的焦点放在这些路由协议的理解上，还有它们针对 IPv6 的更新功能特征。

1. 静态路由

IPv6 中的静态路由以与 IPv4 中相同的方式加以使用和配置。静态路由配置语法与 IPv4 中的相同，例外在于地址是 IPv6 格式的，即“ipv6 route < source > < destination > < distance >”，而对于默认路由是“ipv6 ::/0 < destination > < distance >”。但是，依据 RFC 2461，存在一项特定需求：“一台路由器必须能够确定它的每个邻接路由器的链路本地地址，以便确保一条重定向消息的目标地址邻居路由器的链路本地地址，识别该路由器”。这意味着，不建议使用一个全局单播地址作为下一跳地址，否则 ICMPv6 重定向消息将不能正常工作。

2. RIPng（用于 IPv6 的 RIP）

RIP 是一种 IGRP，它使用距离矢量度量在一个 AS 内进行路由。换句话说，不管特定跳或链路的状态为何，路由的成本是以路由中的跳数度量的。RIP 所允许的最大跳数是 15。这个跳限制防止路由环路，也限制了 RIP 可支持的网络大小。RIP 实现水平分割、路由毒化和抑制机制，防止不正确的路由信息进行传播。在一个 AS 中的每台路由器都有运行在其 UDP 进程上的一个 RIP 进程。为在网络间交换路由信息，RIP 仅有两种类型的消息：请求和响应，这是在 UDP 数据报中传输的。每隔 30s，每个 RIP 路由器将更新作为非请求的响应消息传输到每个邻接路由器。这些更新被发送到保留的 UDP 端口号 520。随着网络在尺寸上的增长，将导致每隔 30s 的海量流量突发，这间接地对 RIP 所支持的网络尺寸施加了限制。在多数当前的联网环境中，RIP 不是路由的优先选择，这是因为其收敛时间和扩展性都不佳导致的。RIP 的主要优势是它可在没有任何复杂性的情况下采用小的代码尺寸，容易地加以实现，这使之对小型网络节点、内嵌系统等是有用的。

在 RFC 1058 中定义了带有分类型路由输入的原始 RIP 版本 1（RIPv1）。周期性的路由更新没有携带子网信息，这样缺乏对可变长度子网掩码（VLSM）的支持。这个限制不支持在一个网络类内有不同大小的子网。不出差错的安全措施的缺乏使 RIPv4 对各种攻击是脆弱的。在 RFC 2453 中开发了 RIP 版本 2（RIPv2），使之变得更加鲁棒。RIPv2 将一些额外的信息添加到路由信息，包括一些安全考虑。RIPv2 包括携带子网信息的能力，因此支持 CIDR。它有必要的后向兼容特征，以便与 RIPv1 一起正常工作。在 RIPv2 中添加路由标签，将来自内部路由的路由与来自 EGRP 的外部重分配（Redistributed）路由区分开来。为约束路由信息分组洪泛网络，RIPv2 组播路由信息分组被发送到处在地址 224.0.0.9 的邻接路由器。作为一项安全措施，RIPv2 使用 MD-5 认证进行路由认证。

在 RFC 2080 中定义了 RIPng，作为支持 IPv6 的 RIPv2 的一个扩展。设计 RIPng，尽可能地与 RIPv2 相似。事实上，相比 RIPv2，RIPng 不引入任何特定的新特征，例外是那些需要在 IPv6 上实现的特征，以及没有去除跳距离的限制。这是在维护 IPv6 简单性需要的情况下完成的，从而使之可实现在非常简单的设备上，而在这些设备上实现 OSPFv3 将会出现问题。下面是实现在 RIPng 中对 RIPv2 的 5

个特定扩展：

1）无类寻址支持和子网掩码指定。在 IPv6 中，所有地址都是无类的，并使用一个地址和一个前缀长度而不是子网掩码加以指定。由此，为每项提供前缀长度的一个字段，而不是一个子网掩码字段。

2）下一跳指定。在 RIPng 中维护这个特征，但以不同方式实现。因为它是一个可选特征，只有在需要时，才在一个独立的路由表项中指定。

3）认证。RIPng 没有其自己的认证机制。认证和加密是作为因特网协议安全（IPsec）特征（在 IP 层为 IPv6 定义的）组成部分实现的。

4）路由标签。这个字段是以 RIPv2 中相同的方式实现的。

5）组播的使用。为 RIP 更新，RIPng 使用组播地址 FF02::9。

存在两个基本的 RIPng 消息类型，即请求和应答，它们是使用 UDP 数据报交换的。RIPng 路由更新消息被发送到著名的端口号 521。RIPng 的消息格式类似于 RIPv2 的消息格式，例外是路由表项的格式不同。图 2.45 给出了 RIPv6 分组格式。

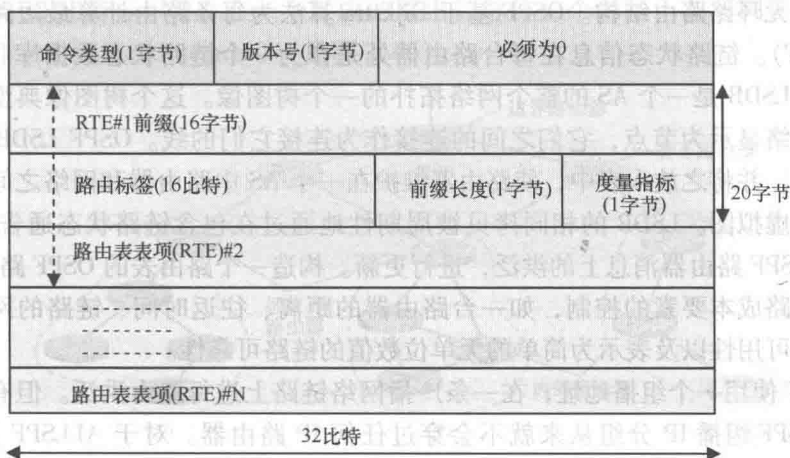


图 2.45　RIPng 分组格式

命令类型字段是 1 个字节，并识别要发送的 RIPng 消息的类型：1 用于一条 RIPng 请求，2 用于一条 RIPng 响应。这些命令被用在两个邻居之间的路由表。版本号字段指定当前版本，即 1。必须为零（Must be Zero）字段是一个保留字段，值必须被设置为 0。路由表项（RTE）由 20 字节组成。每个 RTE 包含一个目的 IPv6 前缀（16 字节）、一个前缀长度（1 字节）、一个路由标签（2 字节）（提供这条路由要携带的额外信息）和一个度量（指定到达目的网络的跳距离（1～15））。当需要指定下一跳时，在带有路由标签且前缀长度 =0 以及度量字段设置为 255（0xFF）的所有 RTE 之前，包括一个特殊的 RTE #1。

3. OSPFv3

OSPF 是一种 IGRP，它仅在单个 AS 内路由 IP 分组。它是 IP 网络的一个自适

应路由协议，并使用链路状态算法和 Dijkstra 最低成本路径算法。它最初是针对 IPv4 在 RFC 2328（OSPFv2）中定义的，并在 RFC 2740[47] 中扩展支持 IPv6，且它最终在 RFC 5340[57] 中成为 OSPFv3 的一个独立标准。在 OSPFv2 和 OSPFv3 中存在一些高度的相似性。虽然 OSPFv3 使用与 OSPFv2 相同的基本机制，但它不是后向兼容于 OSPFv2（如果希望使用 OSPF 在 IPv4 和 IPv6 中路由，则必须同时运行 OSPFv2 和 OSPFv3）。本节开始对 OSPFv2 和 OSPFv3 的共同基础机制的讨论，之后继续讨论理解这两者之间的特定操作方面的差异。

（1）OSPF 基础

OSPF 是一个 AS 内部的协议。OSPF 是一种链路本地协议，它使用链路状态信息的洪泛和 Dijkstra 的最小成本路径算法。OSPF 使用从路由器得到的链路状态信息，构造网络的一个拓扑图。这个图确定 IP 层的路由表，它基于目的 IP 地址做出路由决策。由此，OSPF 没有使用 TCP 或 UDP，而是直接封装在一条 IP 数据报中。OSPF 支持 CIDR 地址模型，并非常快速地检测拓扑中的变化，并在数秒内收敛到一个新的无环路路由结构。OSPF 基于 Dijkstra 算法为每条路由计算最短路径优先树（SPFT）。链路状态信息在每台路由器处是作为一个链路状态数据库（LSDB）维护的。LSDB 是一个 AS 的整个网络拓扑的一个树图像。这个树图像典型地将路由器和网络显示为节点，它们之间的连接作为连接它们的线。OSPF LSDB 实现取这个信息，并将之放入表中，使路由器维护在一个 AS 中路由器和网络之间所有连接的一个虚拟图。LSDB 的相同拷贝被周期性地通过在包含链路状态通告（LSA）的所有 OSPF 路由器消息上的洪泛，进行更新。构造一个路由表的 OSPF 路由策略，也受到链路成本要素的控制，如一台路由器的距离、往返时间、链路的网络吞吐量、链路可用性以及表示为简单的无单位数值的链路可靠性。

OSPF 使用一个组播地址，在一条广播网络链路上进行路由洪泛。但有趣的是指出，OSPF 组播 IP 分组从来就不会穿过任何 IP 路由器。对于 ALLSPF 路由器，OSPF 为 IPv4 保留组播地址 224.0.0.5，为 IPv6 保留 FF02::5；对于所有指定路由器，为 IPv4 保留 224.0.0.6，为 IPv6 保留 FF02::6。当运行在 IPv4 上时，OSPF 协议可使用各种认证方法安全地运作，而运行在 IPv6 上时，OSPFv3 依赖于 IPsec 实现。在 IPv4 上 OSPF 基于子网运行，而在 IPv6 上，OSPFv3 依据链路而运行。无论何时在一条链路的状态发生改变时，一台路由器广播链路状态信息。至少在每 30 分钟内，该路由器也周期性地广播一条链路状态。

OSPF 被设计为有利于较小型和较大型 AS 中的路由。到此为止，该协议支持两种类型的拓扑，即扁平的和层次结构的。

（2）OSPF 扁平拓扑

对于具有少量路由器的一个 AS，整个 AS 可作为单个扁平实体加以管理。通过使用 LSA 而与路由器的对端通信，每台路由器维护一个等同的 LSDB。通过每隔 30 分钟的 LSA 例程性地更新，保持所有 LSDB 是同步的和最新的。同样没有多少信息

需要到处发送，原因是 AS 较小。这种比较简单的拓扑确实扩展起来非常好，并可支持许多较小型的、甚至中等尺寸的 AS。但是，随着路由器数量增加，更新 LSDB 所需的通信也将增加。由此，在具有数百台路由器的一个大型互联网络中，使用一个扁平拓扑，使所有路由器均为 OSPF 对端，可能导致性能降级。这是由于带有路由信息分组的网络洪泛和大型 LSDB 的维护，其中包括整个 AS 中每台路由器和网络。

(3) OSPF 层次结构拓扑

OSPF 支持使用一种层次结构型拓扑，为较大型的互联网络提供较佳的支持。在这种情形中，AS 不再被看作互联路由器的单个扁平结构，其中所有路由器都是对端。相反会构造一个两层层次结构拓扑。但是，AS 仍然是该层次结构的根。一个 OSPF AS 被细分为区，每个区包含一组互联网络。这些区被标记一个 32 比特的区标识符，是以点分十进制格式写出的（如 w. x. y. z，但它们不是 IP 地址）。所以每个区几乎就像它本身就是一个 AS。IPv6 的 OSPF，即 OSPFv3，也以相同表示使用同样的 32 比特标识符。在一个区内的路由被称作一个内区，在不同区之间的路由被称作一个间区（Interarea）（见图 2. 46）。

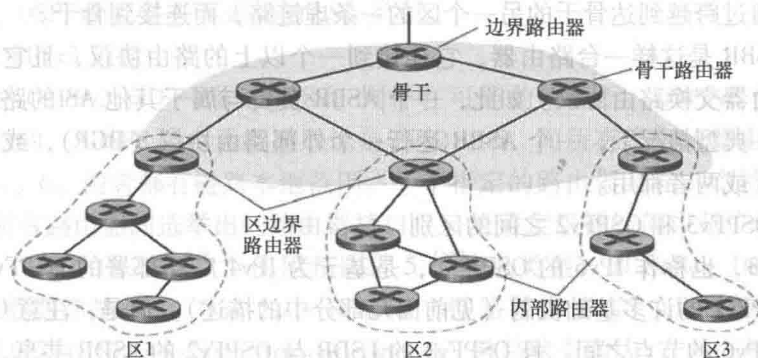


图 2. 46　带有四个区的层次结构型 OSPF AS

每个区运行其自己的 OSPF 链路状态算法，其中每台路由器在一个区中广播它包含在 LSA 的链路状态信息到该区中的所有其他路由器。在任何区中的路由器维护一个 LSDB，它包含有关那个区内各路由器和各网络的信息。在每个区内，一台或多台路由器被配置为区边界路由器（ABR），它负责路由分组到该区外部。连接一个以上区的这些路由器，维护它们作为组成部分的每个区的 LSDB，也将各区联系在一起以便在其间共享路由信息。准确地说，在 AS 中的一个 OSPF 区被配置为骨干区，也称作区 0 或区 0. 0. 0. 0（见图 2. 46）。所有其他区被连接到骨干区。骨干区的主要角色是在 AS 的其他区之间路由流量。骨干区总是包含所有区边界路由器，也可能包含非边界路由器。通过连接到骨干区和到其自己关联区的路由器，发生间区路由。

一般而言，OSPF 定义四种类型的路由器：内部路由器（IR）、区边界路由器（ABR）、自治系统边界路由器（ASBR）和骨干路由器（BR）（见图2.46）。每台路由器有定制地写为点分十进制格式（w. x. y. z）的一个标识符。这个标识符必须在每个 OSPF 实例中建立。但是，因为这个标识符与IP地址没有任何关系，所以就不需要是网络中任何可路由子网的组成部分。重要的是指出，路由器类型是OSPF进程的一个属性。一个给定的物理路由器可有一个或多个 OSPF 进程。例如，连接到一个以上区并从 BGP 进程接收路由信息的一台路由器，是一个 ABR 也是一个 ASBR。

1）IR是这样一台路由器，它在同一区中的各接口具有 OSPF 邻居关系，即一个 IR 在单个区内有其所有的接口。由此，它仅有 OSPF 进程的单个实例。

2）ABR 是这样一台路由器，它将骨干连接到一个或多个区，并有 OSPF 进程的多个实例，一个区一个实例，骨干有一个实例。它在内存中也保有 LSDB 的多个拷贝，对于那台路由器连接到的每个区有一个拷贝。

3）BR 是这样一台路由器，它与骨干区有一个接口，不管它是一台边界路由器还是骨干区的一个 IR。一个 ABR 总是一个 BR，原因是所有区必须直接连接到骨干，或通过跨越到达骨干的另一个区的一条虚链路，而连接到骨干。

4）ASBR 是这样一台路由器，它连接到一个以上的路由协议，且它与其他协议中的路由器交换路由协议。如此，一个 ASBR 负责与属于其他 AS 的路由器交换路由信息。典型情况下，一个 ASBR 运行一个外部路由协议（BGP），或使用一条静态路由，或两者都用。

（4）OSPFv3 和 OSPFv2 之间的区别

OSPFv3，也称作 IPv6 的 OSPF[56]，是基于为 IPv4 广泛部署的OSPFv2 的，且维持了 OSPFv2 的许多基础机制（见前面几部分中的描述）。但是，注意 OSPFv3 运行在支持 IPv6 的节点之间，且 OSPFv3 的 LSDB 与 OSPFv2 的 LSDB 共享。OSPF 的两个版本将并行地操作。OSPFv3 直接层叠在 IPv6 之上，OSPF 首部是由前一首部的下一首部字段中的值 89 识别的。在 OSPFv3 中，对规程和 LSA 存在几项改变。将在下面描述一些重要改变：

1）每链路协议处理：在 IPv6 中，到一条链路的一个接口可有一个以上的 IP 地址。换句话说，单条链路可属于 IPv6 中的多个子网，附接到相同链路但属于不同子网的两个接口可进行通信。考虑这个事实，OSPFv3 支持在相同链路上但属于不同 IPv6 子网的两个邻居进行分组交换。

2）去除地址语义：在 OSPFv3 中，路由器和网络 LSA 不携带 IP 地址。为那个目的定义了一条新的 LSA，它有一些扩展优势。但是，在 IPv6 中维护一个 32 比特的 RouterID（RID）、AreaID（AID）和 LSA-ID。

3）各邻居总是通过 RID 识别的：在 OSPFv3 中，在所有链路类型上的所有邻居都可由 RID 加以识别。

4）添加链路本地洪泛范围：OSPFv3 在保留 OSPFv2 的 AS 或域和区洪泛范围时，也添加了一个链路本地洪泛范围。为携带仅与单条链路上各邻居有关的信息，添加了新的 LSA，称为链路 LSA。这是链路本地范围，即它不能被洪泛超出任何附接的路由器。

5）链路本地地址的使用：OSPFv3 使用路由器链路本地 IPv6 地址（总是以 FF80::/10 开始）作为源地址和下一跳地址。相比而言，OSPFv2 分组有一个本地链路范围，所以它们不被转发到任何路由器。

6）对一条链路多个实例的支持：为区别各实例，通过向 OSPF 分组首部添加一个实例 ID，OSPFv3 支持每条链路多个实例。这项设施用于多台 OSPF 路由器可被附接到单条广播链路［如共享的网络接入点（NAP）］的那些应用。

7）去除 OSPF 特定的认证：IPv6 有其自己的标准认证规程 IPsec，使用认证扩展首部。所以，OSPF 不需要有其自己的认证。

8）未知 LSA 的更灵活处理：OSPFv3 可将其处理为具有链路本地范围的洪泛，或就像它们可被理解一样进行存储和转发，而在其自己的 SPF 算法中忽略它们。相比 OSPFv2，这项设施在比较容易的网络改变和新能力的比较容易集成方面，可有所助益。OSPFv2 总是丢弃未知的 LSA 类型。

（5）OSPFv3 消息

IPv6 的 OSPFv3 有下一首部值 89。同样在可能的情况下，OSPFv3 使用组播。如前面指出的，AllSPF 路由器组播地址是 FF02::5，而所有指定的路由器组播地址是 FF02::6。两者都有链路本地范围。一个指定的路由器是在一个特定的组播访问段上所有路由器间选举出的路由器接口，一般假定组播访问段是广播多址的。与 OSPFv2 一样，OSPFv3[48,57] 使用相同的 5 个消息类型，即 Hello、Database Description（数据库描述）、Link-State Request（链路状态请求）、Link-State Update（链路状态更新）和 Link-State Acknowledgment（链路状态确认），并以相同方式对这些消息进行编号。除了类型字段不同外，对所有类型的消息，消息首部是相同的（见图 2.47）。

各字段的使用和含义如下：

1）版本号：对于 OSPFv3，设置为 3。

2）类型值：Hello 是 1，Database Description 是 2，Link-State Request 是 3，Link-State Update 是 4，Link-State Acknowledgment 是 5。

3）分组长度：是以字节表示的消息长度，包括这个首部的 16 字节。

4）路由器 ID：是产生这条消息的路由器的 ID。

5）区 ID：是这条消息所属的 OSPF 区的一个 32 比特标识符。

6）实例 ID：支持 OSPF 的多个实例运行在单条链路上。没有认证字段。

Hello 消息是由路由器使用的，用来发现在其本地链路和网络上的其他邻接路由器。这条消息在邻接设备之间建立一种关系，称作邻接关系（Adjacencies），并

版本号(1字节)	类型(1字节)	分组长度
路由器ID		
区ID		
检验和(2字节) 0		实例ID(1字节)

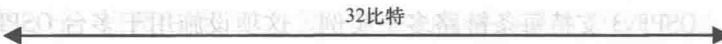


图 2.47　OSPFv3 分组首部

就 OSPF 如何在 AS 中使用沟通关键参数。不像 OSPFv2 的是，没有网络掩码字段，因为 IPv6 不需要这个字段。但 OSPFv3 的选项字段增加到 24 比特，而路由器死亡间隔字段从 32 比特减少到 16 比特。整个选项的情况见 OSPFv3（见图 2.48）。

Hello分组	
类型=1的消息首部	
接口ID	
路由器优先级(1字节)	选项
Hello间隔(2字节)	路由器死亡间隔(2字节)
指定的路由器ID	
备份的指定路由器ID	
邻居ID#1	

邻居ID#N	

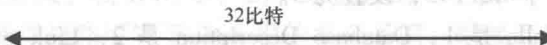


图 2.48　OSPFv3 hello 消息格式

各字段的使用和含义如下：

1）接口 ID：指一个 32 比特数，在这台路由器的接口集间唯一地标识这个接口。

2）Hello 间隔：指在发送 hello 消息之间，这台路由器要等待的秒数。

3）选项：指明该路由器支持哪些可选的 OSPF 能力。

4）路由器优先级：指明当选取备份指定路由器时这台路由器的优先级。

5）路由器死亡间隔：在路由器被认为死亡之前它可保持沉默的秒数。

6）指定的路由器 ID：指明对于一些网络上的某些特定功能的指定路由器地址，且如果没有指定路由器，则被设置为 0。

7）备份的指定路由器 ID：指明备份指定路由器的地址，且如果没有备份指定路由器，则被设置为 0。

8）邻居 ID：指明最近这条路由器接收到 hello 消息的每台路由器的地址。

数据库描述消息从一台路由器将一个 AS 或一个区的 LSDB 内容传递给另一台路由器。传递一个大型的 LSDB，要求发送几条消息。这是采用如下方法完成的，使发送节点指定为主设备，并按顺序发送消息，其中 LSDB 信息的接收者指定为从设备，以确认做出响应。这条消息仅在大型选项字段（见图 2. 49）才与 OSPFv2 的相应消息有所区分。

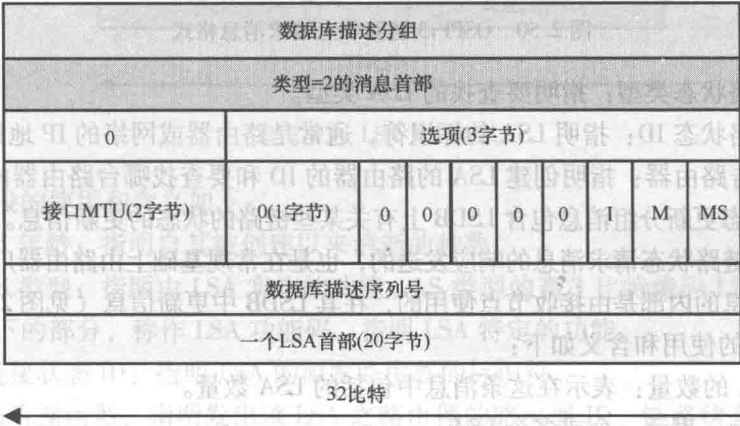


图 2. 49　OSPFv3 数据库描述分组格式

这些字段的使用和含义如下：

1）接口 MTU：指明在没有分片的情况下，在这台路由器的接口上发送的最大消息的尺寸。

2）标志：I 比特被设置为 1，指明这是 DD 消息序列中的第一条消息；M 比特被设置为 1，指明更多的 DD 消息后跟这条消息；MS 比特被设置为 1（如果发送这条消息的路由器是通信中的主控方）或为 0（如果它是从属方）。

3）数据库描述序列号：指明 DD 消息的序列号。LSA 字段包含 LSA 首部（在下面各节做了解释），它携带有关 LSDB 的信息。

链路状态请求消息消息由一台路由器使用，从另一台路由器请求有关 LSDB 一部分的更新信息。该消息准确地指明要查找那条链路的哪种当前信息。这个消息与 OSPFv2 相应消息相同（见图 2. 50）。

这些字段的使用和含义如下：

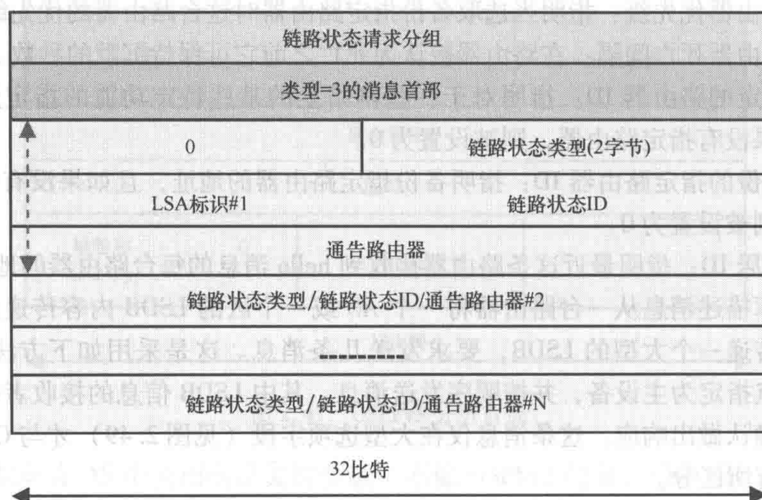


图 2.50　OSPFv3 链路状态请求消息格式

1）链路状态类型：指明要查找的 LSA 类型。

2）链路状态 ID：指明 LSA 的标识符，通常是路由器或网络的 IP 地址。

3）通告路由器：指明创建 LSA 的路由器的 ID 和要查找哪台路由器的更新。

链路状态更新分组消息包含 LSDB 上有关某些链路的状态的更新信息。这条消息是作为一条链路状态请求消息的响应发送的，也是在常规基础上由路由器广播或组播的。这条消息的内部是由接收节点使用的，在其 LSDB 中更新信息（见图 2.51）。

各字段的使用和含义如下：

1）LSA 的数量：表示在这条消息中包括的 LSA 数量。

2）LSAs：表示一个或多个 LSA。

通过显式地确认一条链路状态更新消息的接收，这条消息向链路状态交换过程提供可靠性。LSA 首部识别确认的 LSA（见图 2.52）。

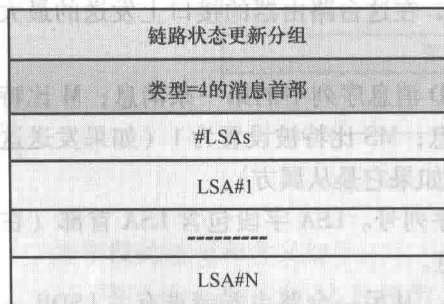


图 2.51　OSPFv3 链路状态更新分组格式

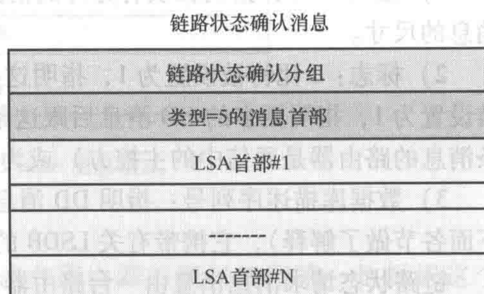


图 2.52　OSPFv3 链路状态确认分组格式

(6) 链路状态通告

注意，上述消息中的几条消息包括多个 LSA。LSA 是携带有关 LSDB 的拓扑信息的字段。存在几种类型的 LSA，它们被用来携带有关不同链路类型的信息。像 OSPF 消息一样，每个 LSA 都有一个 20 字节的通用首部，之后是附加字段（描述一个特定 LSA）的数量。图 2.53 给出了 LSA 首部格式。这几乎等同于 OSPFv2 LSA 格式，例外是没有选项字段，而 LS 类型字段是 16 比特的。

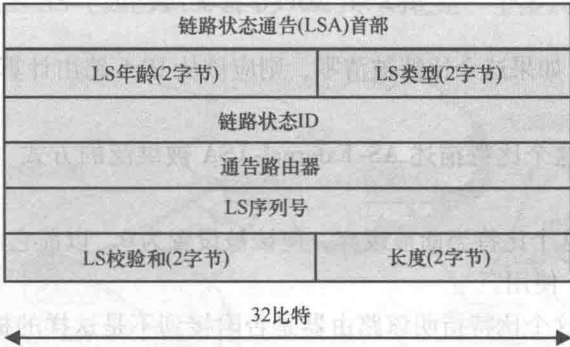


图 2.53 LSA 首部格式

各字段的使用和含义如下：

1）LS 年龄：指明自其被创建以来消逝的秒数。

2）LS 类型：指明由 LSA 实施的功能。LS 类型的高 3 比特编码 LSA 的通用性质，而剩下的部分，称作 LSA 功能码，指明 LSA 特定的功能。

3）链接状态 ID：指明 LSA 的源发路由器的标识符。

4）通告路由器：指明发出该 LSA 之路由器的路由器 ID。链路状态 ID、LS 类型和通告路由器唯一地识别 LSDB 中的 LSA。

5）LS 序列号：指明一个 LSA 后续实例的序列号，被用来检测一个 LSA 的重复实例。

6）LS 校验和：是 LSA 的一个 Fletcher 校验和，用于数据损坏保护（corruption protection）。

7）长度：以字节数表示的 LSA 长度，包括一个 20 字节的 LSA 首部。

下面的 LSA 首部是一个特定 LSA 的特定字段，唯一地由链路状态 ID、LS 类型和通告路由器字段的组合体加以识别。在参考文献中定义了几种类型的 LSA，用来传递有关不同类型链路的信息。所有这些 LSA 的讨论超出了本章的范围。欲了解细节，请参见 RFC 5340[56]。

(7) OSPFv3 选项字段

在 hello 分组、数据库描述分组和某些类型的 LSA［路由器 LSA、网络 LSA、区间（inter-area）路由器 LSA 和链路 LSA］中存在 24 比特的 OSPF 选项字段。选

项字段使 OSPF 路由器支持或不支持可选的能力，并将它们的能力等级传递到其他路由器。通过这种机制，不同能力的路由器可在一个 OSPF 路由域内混合使用。仅指派了 7 比特的 OSPF 选项字段。下面简短地描述每个比特（见图 2.54）。

比特	0	1	2	3	4	5	6	7	8	9	10	12	13	14	15	16	17	18	19	20	21	22	23
																*	*	DC	R	N	*	E	V6

图 2.54 OSPFv3 选项字段

1）V6 比特：如果这个比特被清零，则应该从 IPv6 路由计算中排除该路由器/链路。

2）E 比特：这个比特描述 AS-External-LSA 被洪泛的方式（其解释超出本章的范围）。

3）x 比特：这个比特当前被废弃，应该被设置为 0。以前它由组播开放最短路径优先（MOSPF）使用。

4）N 比特：这个比特指明该路由器是否附接到不是这样的桩区（not-so-stubby area）（NSSA）。一个桩区是这样的一个区，它不从一个 AS 外部接收路由通告，在该区内的路由是完全基于一条默认路由的。NSSA 是一种类型的桩区，它可输入 AS 外部路由，并将它们发送到其他区，但不能为其他区接收 AS 外部路由。桩区和 NSSA 的更多细节超出了本章的范围。

5）R 比特：指明源发方是否为一台活跃的路由器。如果这个比特为零，则不能计算通告节点的路由器。

6）DC 比特：描述应需电路的路由器处理。

7）* 比特：这些是为 OSPFv2 协议扩展的迁移保留的。

4. 边界网关协议版本 4

BGP 是一种 EGRP，通过在现代 TCP/IP 互联网络（被流行地称作因特网）的多个 AS 之间交换路由和可达性信息而实施路由。最初是在 20 世纪 80 年代后期作为 EGP 的一个后继者开发的，BGP 已经修改了多次；在 RFC 4271（1771）[42,54] 中规范当前版本 BGP4。BGP4 支持 AS 的一个任意拓扑。每个 AS 指派一台或多台路由器来实现 BGP。这些路由器接下来交换信息，相互建立联系，并使用 TCP 通过因特网共享有关路由的信息。BGP4 是一种路径矢量路由协议，携带所穿越路由的路径信息（属性）。BGP4 也是一种无类路由协议，它使用 CIDR 地址表示，而不管 AS 是在运行有类的还是无类的 IGRP。

（1）BGP4 操作

配置使用 BGP4 的每台路由器被称作 BGP4 代言器（Speaker）。各设备使用 BGP4 消息通信系统（在下面各节中解释）交换路由信息。例如，在图 2.55 中，路由器 Rij 是 BGP 代言器。仅连接到相同 AS 中其他代言器的 BGP 代言器被称作 IR

(如图 2.55 中的路由器 R1C)，而连接到其他 AS 的那些 BGP 代言器被称作边界路由器（例如，图 2.55 中的路由器 R1A、R1B、R2A、R2B 和 R3A)。在相同 AS 中的邻接 BGP 代言器被称作内部对端（如图 2.55 中的 R1B 和 R1C)，而在不同 AS 中的那些 BGP 代言器被称作外部对端（如图 2.55 中的 R1A 和 R2A)。同样，在一个 BGP4 互联网络中的每个 AS 是一个桩 AS 或一个多穴连接 AS。如果一个 AS 仅连接到另外一个 AS（如图 2.55 中的 AS1 和 AS3)，则它就是一个桩 AS，如果它连接到两个或多个其他的 AS（如图 2.55 中的 AS2)，则它是一个多穴连接的 AS。

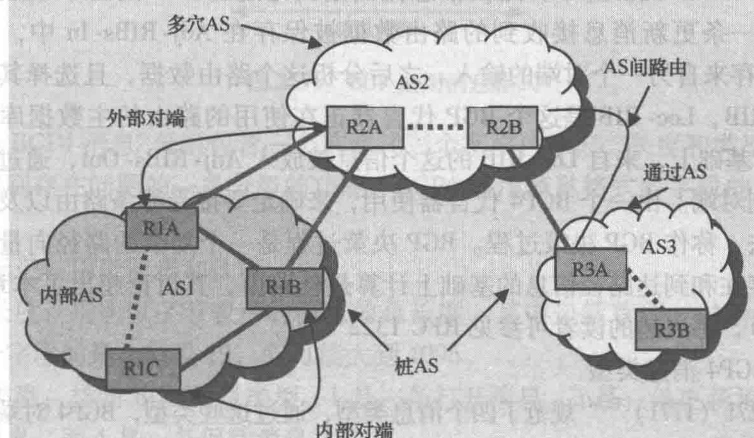


图 2.55　BGP 拓扑图示

BGP 实施三种类型的路由：AS 间路由、AS 内路由和通过（pass-through）AS 路由。AS 间系统路由发生在不同 AS 中的两个或多个 BGP 代言器之间（如图 2.55 中的 R1A 和 R2A)。AS 内系统路由发生在位于相同 AS 内的两个或多个 BGP 代言器之间（如图 2.55 中的 R1A 和 R1C)。通过 AS 路由发生在这样的两个或多个 BGP 对端代言器之间，它们通过不运行 BGP 的一个 AS 交换流量。这种类型的路由一般发生在两个 AS 通过一个多穴连接的 AS 连接时的情况（图 2.55 中的 AS1 到 AS2 到 AS3)。在这种情形中，BGP 允许多穴连接 AS 的管理员建立路由策略，这些策略指定在什么条件下多穴连接 AS 愿意处理中转流量（在一个多穴连接 AS 之上发送的流量)，这些流量的源和目的地都在多穴连接 AS 的外部。

(2) BGP4 路由

BGP4，按照任何路由协议的要求，也维护路由信息，传输路由更新，并基于这个信息做出路由决策。由此，一个 BGP4 系统的主要功能是与另一个 BGP4 系统交换网络可达性信息，包括有关 AS 路径列表的信息。这个信息被用来构造 AS 连通性的一个图，由该图做出路由决策。这样 BGP 的路由操作要求 BGP 代言器存储、更新、选择和通告路由信息。在 BGP4 中为这个目的使用的中心数据结构是路由信息库（RIB)。RIB 由三节组成：输入数据库的一个集合 Adj-RIBs-In，保持从

对端接收到的信息；一个本地数据库（Loc-RIB），包含路由器的当前路由；输出数据库的一个集合 Adj-RIBs-out，由路由器/代言器使用，将其路由信息发往其他路由器/代言器。此外，RIB 可被实现为有一个内部结构（该结构表示不同组件）的单个数据库或实现为独立的数据库。如在前面指出的，BGP4 是一种路径矢量协议，路由在 BGP4 中被称作路径。对于路由，BGP 不仅存储到达目的地的单纯路径，而且以 BGP4 路径属性的形式存储路径的一个详细描述，这使 RIB 成为一个十分复杂的数据结构。当研究消息格式时，将讨论这些路径属性。

在一个 BGP4 代言器中，路由信息流由刚讨论过的 RIB 的三节组成。由对端 BGP4 通过一条更新消息接收到的路由数据被保存在 Adj-RIBs-In 中，每个 Adj-RIBs-In 保存来自另一个对端的输入。之后分析这个路由数据，且选择其合适部分更新 Loc-RIB，Loc-RIB 是这个 BGP 代言器正在使用的路由的主数据库。在周期（Regular）基础上，来自 Loc-RIB 的这个信息被放入 Adj-RIBs-Out，通过一条更新消息发送到对端。由一个 BGP4 代言器使用，来确定要接受哪条路由以及通告哪条路由的方法，称作 BGP 决策过程。BGP 决策过程是一个复杂的路径向量算法，该算法在预存在和到达路径信息的基础上计算最佳路由。其讨论超出了本章的范围。欲了解细节，感兴趣的读者可参见 RFC 1322[41]。

（3）BGP4 消息类型

RFC 4271 （1771）[42,54]规范了四个消息类型，通过这些类型，BGP4 对端设备（代言器）进行通信：打开（Open）消息、更新（Update）消息、通知（Notification）消息和保活（Keep-Alive）消息。

打开消息在 BGP 对端设备之间打开一个 BGP4 通信会话，是在一条 TCP 连接建立之后由每侧发送的第一条消息。这条消息的目的是在设备之间建立联系，识别消息的发送方及其 AS，并协商重要的会话特定的参数。使用保活消息，由对端设备确认打开消息。

为提供路由更新，在对端代言器之间交换一条更新消息。为确保可靠交付，使用 TCP 发送更新消息。更新消息使用一个复杂的结构，这允许一个 BGP 代言器高效地指定新的路由、更新现有路由并撤掉（Withdraw）不再有效的路由。更新消息可从 RIB 撤掉一条或多条不可行的路由，同时在撤掉其他路由时可通告一条路由。

通知消息可用于任意问题报告。每条消息包含一个错误码字段，指明所发生问题的类型。对于某些错误码，一个错误子码字段提供有关问题的特定性质的更多细节。这些消息被用来关闭一个活跃的会话，并通知正被关闭会话的任何连接上的路由器。

保活消息将一个设备是活跃的状况通知 BGP 对端。在空间时段，周期性地发送保活消息，以防止各会话过期。

（4）BGP4 分组格式

这里汇总 BGP4 打开、更新、通知和保活消息以及分组首部格式。注意，BGP

在 TCP 之上发送分组，且 TCP 将数据作为一个字节流发送。所以，BGP 消息可有奇数个字节，且不需要将 BGP 消息分成一个 32 比特或 64 比特的边界。

1）BGP 通用消息格式：所有 BGP4 消息类型使用一个基本的分组首部。除了保活消息外，所有其他消息都有附加字段。一条 BGP4 消息由 4 个字段组成（见图 2.56）。

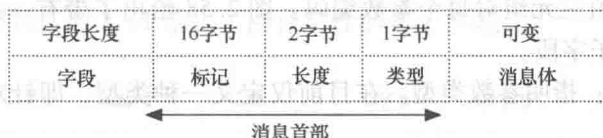


图 2.56 BGP 通用消息格式

每条 BGP4 消息/分组包含一个首部（三个字段：标记、长度和类型），其主要功能是识别存在问题的一条分组的功能。BGP4 通用消息格式各字段的使用和含义如下：

① 标记：用于同步和认证。

② 长度：指明以字节表示的消息的总长度，其中包括首部。对于一条保活消息，这个字段的最小值是 19，它可能大到 4096。

③ 类型：指明 BGP 消息类型，1 是一条打开消息，2 是一条更新消息，3 是一条通知消息，而 4 是一条保活消息。

④ 消息体/数据：包含用于实现打开、更新和通知消息的每个消息类型的各字段。

2）BGP4 打开消息格式：一条 BGP4 打开消息由一个消息首部和附加字段组成。由此，除首部字段外，它由 6 个字段组成（见图 2.57）。

字段长度	19字节	1字节	2字节	2字节	4字节	1字节	可变
字段	首部	版本	AS	保持时间	BGP 标识符	可选参数长度	可选参数

图 2.57 打开消息格式

一条打开消息的各字段的使用和含义如下：

① 版本：指明打开消息发送方正在使用的 BGP 版本。当前值是 4。

② AS：指明打开消息发送方的 AS 号。AS 号和前面讨论的一样，目前是以类似于 IP 地址的一种方式在因特网间以中心方式管理的。

③ 保持时间：指定一个 BGP 对端将允许消息接收之间连接保持静默多少秒。这个值必须至少为 3s。如果它是 0，则指明没有使用保持时间。

④ BGP 标识符：是发送方 BGP 代言器的一个接口的一个 IP 地址，是在启动时

确定的。一旦选择，对于BGP代言器及其所有对端的所有本地接口（LI）都是相同的。

⑤ 可选参数长度：指明可选参数字段的长度，如果存在的话。如果为0，那么在消息中就不存在可选参数。

⑥ 可选参数：包含可选参数的一个列表，如果存在一些参数的话。使用一个标准类型/长度/值三元组对每个参数编码。图2.58给出了带有一条打开消息的可选参数字段的各子字段。

• 参数类型：指明参数类型。在目前仅定义一种类型，即针对认证消息，定义了1。

字段长度	1字节	1字节	可变	—	1字节	1字节	可变
字段	参数类型#1	参数长度#1	参数值#1	—	参数类型#N	参数长度#N	参数值#N

图2.58 一条打开消息的可选参数字段的各子字段

• 参数长度：指明参数值子字段的长度。

• 参数值：提供被传递的参数的值。

3）更新消息格式：这条消息由类型字段设置为2的一个BGP首部和附加字段组成，如图2.59所示。除了首部外，它总共有5个字段。其中，NLRI表示网络层可达性信息。

字段长度	19字节	2字节	可变	2字节	可变	可变
字段	首部	不可行路由的长度	撤销路由	总路径属性长度	路径属性	NLRI

图2.59 BGP更新消息格式

在接收到一条更新消息时，BGP代言器将能够从其RIB添加或删除特定路由，以确保准确性。一条更新消息的各字段的使用和含义如下：

① 不可行路由长度：指明以字节表示的撤销路由（Withdrawn Route）字段的长度。如果为0，则没有路由要撤销，并忽略撤销路由字段。

② 撤销路由：包含这样的IP地址前缀的列表，其路由要从其作用中被撤销。每个地址表示为使用子字段的一个子结构。图2.60给出了一条更新消息的撤销路由字段的各子字段。

字段长度	1字节	可变	—	1字节	可变
字段	长度#1	前缀#1	—	长度#N	前缀#N

图2.60 一条更新消息的撤销路由字段的各子字段

长度子字段指明一个 IP 地址前缀中的比特数，而前缀子字段包含网络（要撤销其路由）的 IP 地址前缀。

③ 总路径属性长度：指明以字节表示的路径属性字段的总长度。如果为 0，则指明在这条消息中不通告路由，所以忽略路径属性和 NLRI 字段。

④ 路径属性：描述被通告路由的路径属性。以一个标准类型/长度/值三元组的形式编码每个属性。图 2.61 给出了路径属性的各子字段。

字段长度	2字节	1或2字节	可变	—	2字节	1或2字节	可变
字段	属性类型 #1	属性长度 #1	属性值 #1	—	属性类型 #N	属性长度 #N	属性值 #N

图 2.61　一条更新消息的路径属性字段的各子字段

● 属性类型：定义属性的类型并对之进行描述。这个字段有一个进一步子结构，后面将做解释。

● 属性长度：指明以字节表示的属性值的长度。正常情况下是 1 字节，但对较长的属性，这要用 2 字节。是如下指明的，在属性类型字段中设置扩展的长度标志。

● 属性值：指明属性的值。这个字段的尺寸和含义取决于路径属性的类型。例如，对于一个 Origin（源发）属性，这是单个整数值，指明一条路径的源发点；对于一条 AS_Path 属性，这个字段包含到该网络的路径中 AS 号的一个可变长度列表。图 2.62 给出了属性类型字段的子字段。

字段长度	1字节					1字节
字段	属性标志					属性类型码
	可选比特	中转比特	部分比特	扩展长度比特	保留(4比特)	

图 2.62　一条更新消息的属性类型字段的各子字段

属性标志字段指定一组标志，这些标志描述属性的性质和如何处理。

为可选属性设置可选比特。

为可选临时（transitive）属性设置临时比特。

设置部分比特，指明在一个可选临时属性上的信息是部分的，且是不确定的。

设置扩展长度比特，指明 2 字节的属性长度。

保留字段是 4 比特，目前所有比特都被设置为 0。

属性类型码字段标识属性类型。属性类型如下：

● Origin（源发）——一个必备属性，指明路径信息的源发者，类型码为 1。

● AS_Path——一个必备属性，指定在路径中涉及的 AS 序列的 AS 号列表，类

型码是2。

• Next_Hop——一个必备属性，指定被用来到达这个目的地的下一跳路由器，类型码是3。

• Multi_Exit_Description——一个可选的和非临时属性。当一条路径包括多个出口或入口点时，使用它的值，4是它的类型码。

• Local_Pref——一个区分（discretionary）属性，用在相同AS中BGP代言器之间的通信中，指明一条特定路由的优先等级。类型码是5。

• Atomic Aggregator（原子汇聚器）——一个区分属性，指明BGP代言器在它所接收到的重叠路由集中选择了不太具体的路由。类型码是6。

• Aggregator（汇聚器）——一个可选的临时属性，包含实施汇聚的路由器的AS号和BGP ID。类型码是7。

• NLRI——包含被通告路由的IP地址前缀的一个列表。使用相同的子结构指定每个地址。图2.63给出了一条更新消息的NLRI子字段的各子字段。其中，长度指明在下面是重要的IP地址前缀字段中的比特数；前缀提供其路由正被通告的网络的IP地址前缀。

字段长度	1字节	可变	—	1字节	可变
字段	长度#1	前缀#1	—	长度#2	前缀#2

图2.63 一条更新消息的NLRI字段的各子字段

4）通知消息格式：这条消息由类型字段设置为3的一个BGP消息首部和附加字段组成，如图2.64所示。除了首部外，这条消息还有3个字段。这条消息或分组被用来向源发路由器的对端指明某种类型的错误条件。

字段长度	19字节	1字节	1字节	可变
字段	首部	错误码	错误子码	错误数据

图2.64 通知消息格式

一条通知消息各字段的使用和含义如下：

① 错误码：指明发生的类型或错误。RFC 4271（1771）[42,54]定义如下错误类型及其代码值：1是一个消息首部错误，2是一个打开消息错误，3是一个更新消息错误，4是保持定时器超时，5是指明一个意外事件的一个有限状态机错误，6是释放，在一台BGP设备的请求下关闭BGP会话。

② 错误子码：提供有关所报告错误性质的更具体信息，细节可在RFC 4271（1771）[41,53]中找到。

③ 错误数据：包含基于错误码和错误子码字段的数据。这个数据用来诊断通知消息的原因。

5）保活消息格式：这条消息由带有类型字段设置为 4 的一个消息首部组成，没有附加字段。这条消息是在对端 BGP 代言器之间周期性交换的，保持它们的 TCP 会话处于活跃状态。作为一个特例，这条消息被用作初始 BGP 会话建立期间一条有效打开消息的确认。

（5）为支持 IPv6 对 BGP4 的多协议扩展

RFC 4760[56] 定义了对 BGP4 的一个扩展，支持为多个网络层协议［如 IPv6、互联网分组交换（IPX）、L3VPN 等］携带路由信息。这些扩展是后向兼容的，即支持这些扩展的一个 BGP 代言器可与不支持这些扩展的 BGP 代言器互操作。为提供后向兼容的多协议能力，RFC 4760 引入两个新的路径属性，即多协议可达的 NLRI（MP_REACH_NLRI）和多协议不可达的 NLRI（MP_UNREACH_NLRI）。属性 MP_REACH_NLRI 被用来携带可达目的地集合以及下一跳信息，用于转发到这些目的地。属性 MP_UNREACH_NLRI 被用来携带不可达目的地的一个集合。这两个属性是可选的和非临时的（Nontransitive）。由此，不支持多协议能力的一个 BGP 代言器将仅忽略在这些属性中携带的信息，不会传递到其他 BGP 代言器。

1）多协议可达的 NLRI：这是一个可选的非临时属性，属性类型码是 14。这个属性可用于两个目的，将一条可行路由通告到一个对端，允许一台路由器将下一跳路由器的网络层地址通告到 MP_REACH_NLRI 属性的 NLRI 字段中列出的各目的地。属性可如图 2.65 所示进行编码。其中，AFI 表示地址族标识符；SAFI 表示后续地址族标识。

字段长度	2字节	1字节	1字节	可变	1字节	可变
字段	AFI	SAFI	下一跳网络地址的长度	下一跳的网络地址	保留	NLRI

图 2.65　MP_REACH_NLRI 属性的编码

这些字段的使用和含义如下：

① AFI：与 SAFI 字段组合，识别在下一跳字段中所承载地址必须属于的网络层协议集、下一跳的地址被编码的方式，以及 NLRI 字段所遵循的语义。

② SAFI：与 AFI 具有完全相同的描述，见 RFC 4760[56]。目前为这个字段定义了两个值：当 NLRI 用于单播转发时为 1，当 NLRI 用于组播转发时为 2。

③ 下一跳网络地址的长度：指明以字节表示的下一跳字段的网络地址的长度。

④ 下一跳的网络地址：包含在到目的地系统的路径上下一跳路由器的网络地址。与这个地址关联的网络层协议是由在属性中携带的元组 < AFI，SAFI > 识别的。

⑤ 保留：必须被设置为 0，并在收到时应该加以忽略。

⑥ NLRI：为在这条属性中通告的可行路由列出 NLRI。NLRI 的语义是由元组 <AFI，SAFI>识别的，并依据 RFC 4760 的 NLRI 编码中规范的进行编码。

注意，携带 MP_REACH_NLRI 的更新消息也必须携带源发（Origin）和 AS_Path 属性。同样，不携带除 MP_REACH_NLRI 属性中编码的其他 NLRI 的一条更新消息，不应携带 Next_Hop 属性。

2）多协议不可达 NRI：是一个可选的非临时属性，属性类型码是 15。它可被用于撤销其作用的多条可行路由的目的。属性 MP_UNREACH_NLRI 如图 2.66 所示进行编码。

字段长度	2字节	1字节	可变
字段	AFI	SAFI	撤销路由NLRI

图 2.66　MP_UNREACH_NLRI 属性的编码

各字段的使用和含义如下：

① AFI：与 MP_REACH_NLRI 属性中给定的描述相同。

② SAFI：与 MP_REACH_NLRI 属性中给定的描述相同。

③ 撤销路由 NLRI：列出正被撤销其作用的路由的 NLRI。NLRI 的语义是由属性中携带的元组 <AFI，SAFI>识别的。NLRI 是按照 RFC 4760 NLRI 编码一节中的规范进行编码的。

注意，不要求包含 MP_UNREACH_NLRI 的一条更新消息携带任何其他的路径属性。

3）NLRI 编码：被编码为形式为 <长度，前缀>的一个或多个元组，如图 2.67 所示。

字段长度	1字节	可变	—	1字节	可变
字段	长度#1	前缀#1	—	长度#N	前缀#N

图 2.67　MP_NLRI 字段的 NLRI 的子字段

各字段的使用和含义如下：

① 长度：指明以比特表示的地址前缀的长度。长度为 0 表示一个前缀匹配所有内容。

② 前缀：包含一个地址前缀，后跟足够的结尾比特（Trailing Bits），使该字段的结束处在一个字节边界上。

4）错误处理：如果一个 BGP 代言器从一个邻居接收到这样一条更新消息，它包含 MP_REACH_NLRI 或 MP_UNREACH_NLRI 属性；如果代言器判定该属性是不正确的，则该代言器必须删除从邻居（其 AFI/SAFI 与不正确的 MP_REACH_NLRI 或 MP_UNREACH_NLRI 中携带的相同）接收到的所有路由。BGP 会话期间（在其

上接收到更新消息），代言器应该忽略在那个会话上接收到 AFI/SAFI 的所有后续路由。此外，BGP 代言器可倾向于终结 BGP 会话，这是在这个会话上接收到更新消息的。在这种情形中，应该以一条带有代码/子代码的通知消息（指明“更新消息错误/可选属性错误”）终止会话。

5）BGP 能力通告：使用多协议扩展的一个 BGP 代言器，应该使用能力通告规程[50]确定代言器可与一个特定对端使用多协议扩展。在能力选项中的字段设置如下：

能力代码字段被设置为 1（指明使用一个多协议扩展）。

能力长度字段被设置为 4。

能力值字段如图 2.68 所示进行设置。

字段长度	2字节	1字节	1字节	—	2字节	1字节	1字节
字段	AFI#1	保留	SAFI#1	—	AFI#N	保留	SAFI#N

图 2.68　能力值字段设置

AFI 和 SAFI 以与多协议扩展相同的方式进行编码。保留字段应该由发送方设置为 0，并由接收方忽略。

注意，为在一对 BGP 代言器之间进行一个特定 <AFI, SAFI> 的路由信息的双向交换，每个这样的代言器必须通告到另一方，这里通过能力通告机制（支持那个特定 <AFI, SAFI> 路由的能力）。

2.5　多穴连接

多穴连接是用来去除因特网连接作为潜在单点故障的一项技术，由此增加一条因特网连接的可靠性。如今，因特网连接对公司和企业网络具有战略重要性，他们希望通过至少两家 ISP 连接到因特网，以增强在提供商网络中出现一个故障时的可靠性，同时改进网络性能。

在 IPv6 语境中，主要存在两种类型的多穴连接现象，即主机多穴连接和站点多穴连接。主机多穴连接可能源自两种情况之一：一台主机有同一范围或不同范围的一个以上的单播地址，或一台主机有一个以上的物理或虚拟接口（见图 2.69）。站点或集群多穴连接也可能来自两种情况之一：一个拓扑集群继承一个以上的地址前缀，或一个站点有到相同或不同前缀的一个以上的外部附接（见图 2.70）。

读者应该注意到主机多穴连接和站点多穴连接之间的并行存在。多穴连接的一台主机必须属于一个多穴连接的站点。同样注意，多穴连接场景未必是静态的。

一台主机可通过主机重新编号协议动态地获取或丢失附加的单播地址，一台主机可通过隧道配置动态地获取或丢失附加的接口，而一个站点可通过路由器重新编号协议动态地获取或丢失附加前缀。特别在 IPv6 中，多穴连接现象是不可避免的。

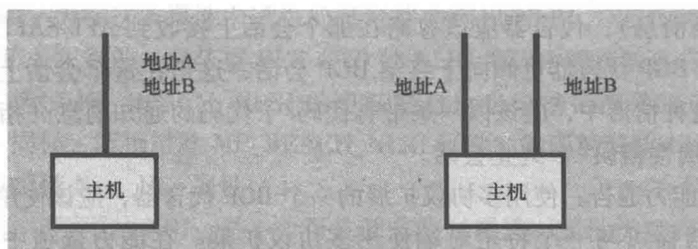


图 2.69 主机多穴连接

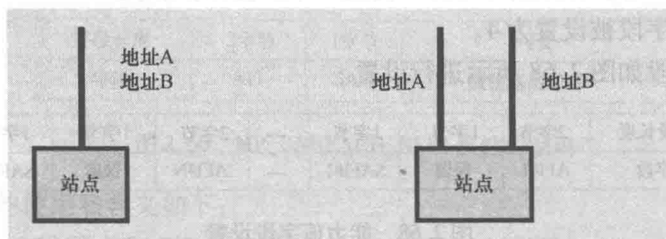


图 2.70 站点多穴连接

首先，IPv6 被设计为针对不同范围和平滑（Graceful）的重新编号，每主机/节点有多个单播地址；其次，人们期望许多主机有多个接口（如无线、有线），而最终附接到多个 ISP 的各站点预计会得到多个前缀。所以，将讨论焦点放在下面各节中以理解这种现象。注意，在下一代 IPv6 中还没有标准化多穴连接。

2.5.1 因特网结构

为理解站点多穴连接，需要就中转路由域（TRD）和端路由域（ERD）方面理解因特网结构。

因特网被组织成路由域，交换有关它所组成的网络可达性的信息。IDRP 将这些路由域细分为 TRD 和 ERD，如图 2.71 所示。

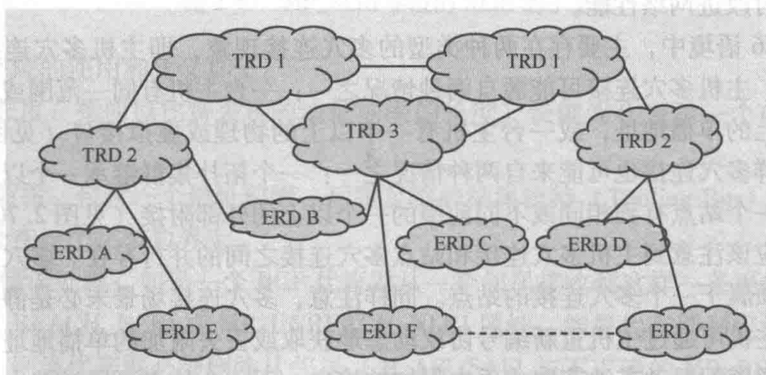


图 2.71 以 TRD 和 ERD 组成的因特网结构

各 ERD 与网络端用户关联，如各组织连接到因特网，同时各 TRD 提供带有中转功能的 ERD。一般而言，每个 ERD 连接到至少一个 TRD。有时，一个 ERD 有与多个 TRD 的多条连接。在这种情形中，ERD 被称作多穴连接（如图 2.71 中的 ERD B）。当两个 ERD 有大体量的流量，而不通过因特网结构时，它们可被连接到一条私有链路。各 TRD 通常与 ISP 关联。这些 ISP 被细分为两类，即直接 ISP［连接到端用户，并将自己连接到一个国际骨干（如美国在线）］和间接 ISP（管理大型国际骨干，是层次结构中的最高层，它们被连接到直接 ISP 或大型用户）。称一个 ERD 为多穴连接的，当在它本身没有成为一个 TRD 情况下，连接到一个以上的 TRD。ERD 多穴连接也被称作站点多穴连接，且 TRD 被直接寻址为 ISP。多穴连接站点的例子有一个大型组织（涵盖整个国家，通过不同 ISP 在多个点连接到因特网）和一个国际组织（将其网络连接到多个国家的因特网，该组织的附属机构位于这些国家）。站点多穴连接的促动因素是防止提供商的链路故障；提供商之间的流量负载均衡；更好的网络性能，包括时延、丢失和抖动；更高的带宽可用性。

2.5.2 主机多穴连接

如前面指出的，主机多穴连接来源于两种状况之一，即有多个单播地址的一台主机或有多个接口的一台主机。无论哪种情况，当处理这样的多穴连接主机时，都需要解决一些问题。

1. 主机多穴连接的问题

(1) 有多个地址的主机多穴连接

在这种情形中，希望发送分组到这样一台多穴连接主机的任意节点，都要选择目的 IP 地址。自然地，发送节点倾向于选择以最短、最快和最廉价的路径可到达意义上效果最好的一个地址。同样，它也可能检测在会话期间一个地址是否停止使用，并能够在应用允许的情况下切换到另一个发挥作用的地址。另外，如果任何应用希望从这样的一个节点发送任意分组，则它要选择源地址。其选择取决于被选中 IP 地址的状况，将影响响应的路径。在一些情形中，如果被选中源地址有不能到达目的地的一个范围，则交付将失败。此外，对于使用地址作为对端标识符的应用来说，我们需要明白，就许多地址识别同一对端方面，这项应用是如何做出结论的。

(2) 带有多个接口的主机多穴连接

在这种情形中，从其他节点到达这样一台多穴连接主机的分组是没有问题的。当从这样一台主机发送分组时，如果源地址不属于外发接口，那么 ICMP 将失败，但为了避免这一点，该主机需要比一台路由器中更复杂的一个路由表。另一个问题是，当在其接口上的源地址失效时，如何处理发送分组。如果人们期望的话，对于到达和外发流量，人们也需要解决多个接口间的负载均衡问题。

2. 主机多穴连接可能解决方案模型

一台多穴连接主机可能有称作本地地址（或LA）的多个地址和称作LI的多个接口。令多穴连接主机希望与之通信的对端，带有称作远端地址（或RA）的多个地址。每个三元组（LA，LI，RA）识别主机和对端之间的一条不同路径。可能的主机多穴连接解决方案模型的范围可做如下考虑：

1）在会话初始时选择一条路径，即最可能发挥作用的一个（LA，LI，RA）三元组。

2）在会话初始时尝试一些或所有的可能路径，直到找到一个可正常工作的路径，如果可能的话，在正常工作的路径间选择最佳路径。

3）将所用路径参数放入缓存，可用于与同一对端的新会话。

4）在一个外发会话中，检测当前路径的失效，并将会话切换到一条更好的正常工作路径。

5）在多条路径间分散会话流量。

同样，以属于源发接口的源地址作用于这些解决方案，将得到不同的主机多穴连接模型。一些IPv6特征将帮助解决一些问题，如一个全局唯一的节点ID，它将为传输协议提供会话期间度过（Survive）地址改变的一种可能，并通过路由重新编号协议（路由重新编号协议超出了本章的范围）控制源地址选择。由此，存在可在非常近的未来就被支持的一些事情，而其他事情会在可预见的未来得到支持。下面讨论参考文献中有关主机多穴连接提出的一些解决方案。

3. IPv6主机多穴连接解决方案

这些解决方案基本上依赖于两件事情，即使用由站点出口路由器支持的多个前缀和增强的主机能力（提供容错、流量工程和路由汇聚[70]）。如此，主机多穴连接解决方案目标是增强主机检测路径故障的能力，并在没有中断传输层会话的任何情况下从一个提供商切换到另一个提供商。由此，在参考文献中提出的这些主机多穴连接解决方案可被粗分为两种主要方法，即传输层方法和网络层方法。本节研究一些这样的方法。为了更好地进行理解，通过描述对所有这些方法都相同的架构，下面开始讨论这些主机多穴连接方法。

（1）架构

采用一个图示，在图2.72中给出，来解释该架构。在这个图示中，两个ISP，即TRD 10和TRD 20，提供到一个多穴连接站点ERD 1120的因特网连接。多穴连接站点ERD 1120从每个提供商接收到一个前缀，即从TRD 10接收到2010：10：1：：/48，从TRD 20接收到2010：20：1：：/48。这两个前缀由站点出口路由器R通告到ERD 1120中的每台主机。这些前缀被用来为每个主机接口依据提供商派生一个IPv6地址。多穴连接站点ERD 1120使用BGP将其前缀2010：10：1：：/48通告到TRD 10，2010：20：1：：/48通告到TRD 20。这些TRD将其自己的IPv6汇聚2010：10：：/32和2010：20：：/32分别通告到全球因特网。为遵守提供

商实施的进入过滤策略，在首部中包含的所有源地址基础上，站点出口路由器选择出口链路。结果，由主机选择的源地址确定所使用的上行（Upstream）提供商。

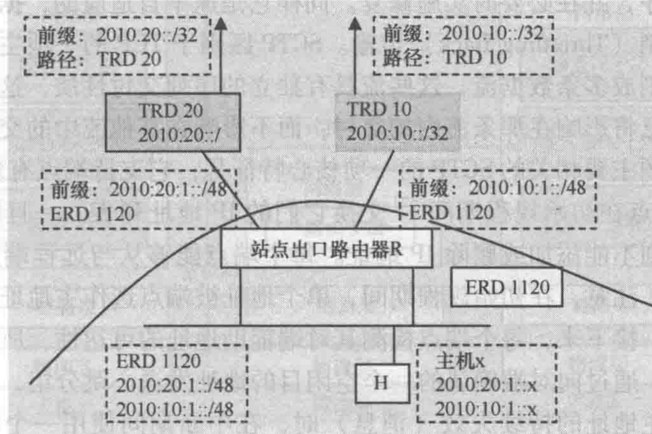


图 2.72 IPv6 主机多穴连接架构

(2) 传输层方法

在这些方法中，多穴连接提供的支持机制位于开放系统互连（OSI）栈的传输层。因为 TCP 和 UDP 的当前传输层协议基于端点的 IP 地址在它们之间建立通信会话，在端点 IP 地址中任何一方的改变都导致会话的中断。如此，它们不处在支持主机多穴连接的位置。所以，这些传输方法建议在传输层中每个端点支持多个地址，从而在不中断通信会话的条件下可以另一个地址替换一个地址。同样，这些机制使用如下事实，即传输层基于分组丢失得到不同路径质量的信息。就安全性而言，基于 cookie 的保护确保每条会话的安全性，但许多安全问题仍然存在。为支持主机多穴连接建议的传输层协议有 TCP-MH、SCTP 和数据报拥塞控制协议（DCCP）。为更好地理解，下面将深入研究这些协议。

1) TCP-MH（TCP 多穴连接）

这是对现有 TCP 的一种扩展，集成了主机多穴连接场景[69]。为支持主机多穴连接，要对 TCP 附加如下内容：

① TCP 的 SYN 分段包含可到达源节点的所有 IP 地址。

② 为在一条已建立的 TCP 连接上通知发送方地址信息的接收者，添加 MH-add 和 MH-delete 选项。在一条连接期间，每次增加或去除可用的 IP 地址时，使用这些选项中的一个。

③ 为了保障 TCP 会话的安全，在 MH-add 和 MH-delete 中添加一个序列号。

无论何时在到另一个可用 IP 地址的一条会话路径中检测到一次中断时，端点使用上述选项。

2）SCTP

流控制传输协议（SCTP）是一条可靠的面向连接单播传输协议，即由其 IP 地址确定的已知端点之间的数据通信。SCTP 提供数据的可靠传输，方法是检测何时数据损坏或乱序，并在必要时实施修复。同样它是速率自适应的，据此对网络拥塞做出响应并抑制（Throttling Back）传输。SCTP 区别于 TCP 的一项主要特征是，它允许数据被分割成多条数据流，这些流具有独立的序列交付性质，这意味着在一条流中丢失的消息将影响在那条流内的交付，而不影响在其他流中的交付。

与这个话题主要相关的 SCTP 的一项核心特征是，它支持端点有多个 IP 地址的能力。SCTP 端点在初始规程期间[62]交换它们的 IP 地址列表。一旦端点之间的关联已经建立，则不能添加或删除 IP 地址。每个端点能够从与远程端点关联的任何端点接收消息。注意，在初始规程期间，单个地址被端点选作主地址，并用作正常传输的目的地。接下来，每个端点检测其对端辅助地址的可达性，所以知道哪些地址可用于切换。通过向对端确认的一个空闲目的地址发送心跳分组，完成监测。当注意到发送到主地址的持续失效（消息）时，在中断期间使用一个辅助地址。直到发送到一个空闲的主地址的心跳分组再次可达之前，使用辅助地址。

3）DCCP

基本上而言，设计 DCCP（数据报拥塞控制协议）是为了控制数据报网络中的拥塞。对 DCCP 的一个扩展[13]提供对移动性的支持，所以支持多穴连接，采用从一个地址到另一个地址的转换连接端点的方式。注意在这种情形中，移动节点必须提前协商这种支持。移动端点得到一个新地址，它将那个地址的 DCCP-move 分组发送到静态端点，站端点将其连接改变到新地址。在这方面，仍然有许多工作要做，在 IPv6 中正在开展这些工作。讨论这些工作超出了本章的范围。

（3）网络层方法

网络层方法通过传输层和网络层之间的一个中间层支持多个地址。这个中间层（称作多穴连接层）的准确位置如图 2.73 所示。如注意到的，多穴连接层位于 IPv6 路由子层（实施网络相关的功能，像分组转发）之上和 IPv6 端点子层（实施端到端功能，像分段和 IPsec）之下。由此，一个新层将包括在一个 IPv6 地址中的两个功能分隔开了：主机的定位符和主机的标识符。主机的定位符指明如何到达主机。它以网络拓扑的方式指定网络附接点。事实上定位符被用来在路由器中转发分组。另外，主机的标识符是在 IPv6 层的一个标签，被呈现给高层（见图 2.73）。事实上，一个标识符被用来将一台主机与另一台主机做出区分，它独立于主机的网络附接。一台主机可有多个标识符，每个标识符都是全局唯一的。注意，IPv6 地址既是定位符又是标识符，原因是它们包含拓扑含义，并作为一个接口的唯一标识符。定位符和标识符的这种隔离，使应用可仅绑定标识符，由多穴连接层将标识符映射到定位符。所以，当一个特定定位符不可用时，标识符被映射到另一个定位符。

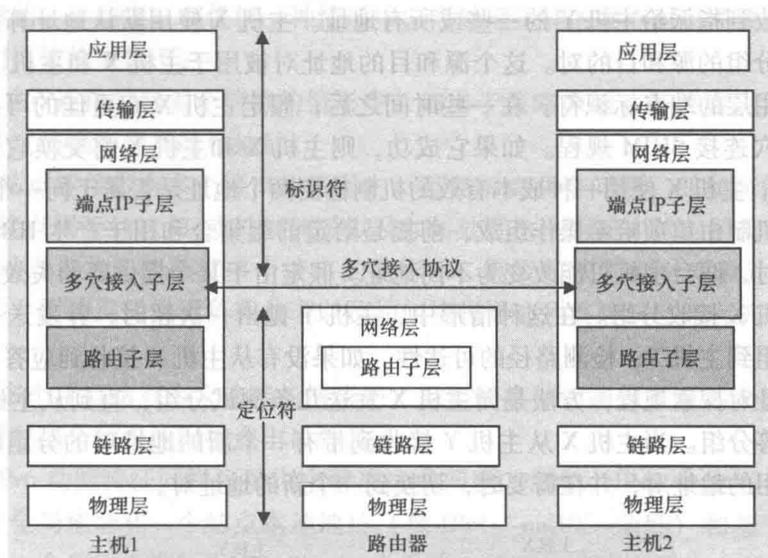


图 2.73　一个协议栈中的多穴连接层

由此，使用这种方法的主机要求称作多穴连接层的一个附加中间层，它具有一致地将呈现给高层的标识符和 IPv6 地址（实际上包含在数据分组中）的机制。在受到攻击时，这种映射是脆弱的，如此，这个协议的任何特定实现都必须考虑安全问题。在参考文献中人们讨论了几种网络层方法。它们主要在实现标识符和定位符之间隔离的实现方法方面存在差异，采取的方法是在网络层和传输层之间定义一个新的中间层。由于篇幅所限，这里不能讨论它们。尽管如此，在下一节，将讨论这些解决方案中最有前景的一种方案，称作 L3 楔子（SHIM）。SHIM 被认为是最著名的，这是由于其高效的安全特征及其对网络基础设施的低要求。

1）SHIM 方法

SHIM[75,79] 方法建议使用位于 IP 路由子层之上和 IP 端点子层之下的一个中间层。这种方法使用可路由的 IPv6 定位符作为标识符，该标识符在 SHIM 层之上是可见的。这些实际上是称作基于哈希的地址（HBA）的一个生成地址集合。这是如下做到的：产生一台主机地址的接口 ID，作为可用前缀和随机数的哈希值。通过将不同网络前缀附加到所产生的接口 ID 之后，产生称作 HBA 的多个地址。作为对影响原定位符失效的响应，在分组地址字段中使用的实际定位符可随时间而改变。

如图 2.74 所示，主机 X 有地址 IP1（X）、IP2（X），主机 Y 有地址 IP1（Y）、IP2（Y）和 IP3（Y）。但传输层和高层可见的稳定源地址和目的地址是 IP1（X）和 IP2（Y）。这里将解释 SHIM 方法的工作过程。方法是描述当一台多穴连接的主机 X 开始与另一台多穴连接主机 Y 发生通信时的事件序列（见图 2.74）。当主机 X 希望发起与主机 Y 的通信时，它首先发起针对主机 Y 的一条 DNS 请求。它在 DNS

响应中接收到指派给主机 Y 的一些或所有地址。主机 X 使用默认地址算法选择将用于外发分组的源和目的对。这个源和目的地址对被用于主机 X 和主机 Y 上所有传输和应用层的端点标识符。在一些时间之后，假定主机 X 有更佳的可靠性，发起一个多穴连接 SHIM 规程。如果它成功，则主机 X 和主机 Y 将交换它们的可用 HBA 集合。主机 X 使用一种成本有效的机制检测两个地址是否属于同一个 HBA 集合。这个机制由单项哈希操作组成，前提是给定前缀集合和用于产生 HBA 的附加参数。此时，两台主机可能改变为不同地址。假定由于某个提供商的失效，主机 Y 不能从主机 X 接收分组。在这种情形中，主机 Y 抛出一次超时，并发送一条可达性测试分组到主机 X，检测路径的可达性。如果没有从主机 X 接收到应答，则它发出一次地址对探索规程，方法是向主机 X 发送几条测试分组，直到从主机 X 接收到一条应答分组。当主机 X 从主机 Y 接收到带有一个新的地址对的分组时，它检查当前使用的地址对，并在需要时，切换到一个新的地址对。

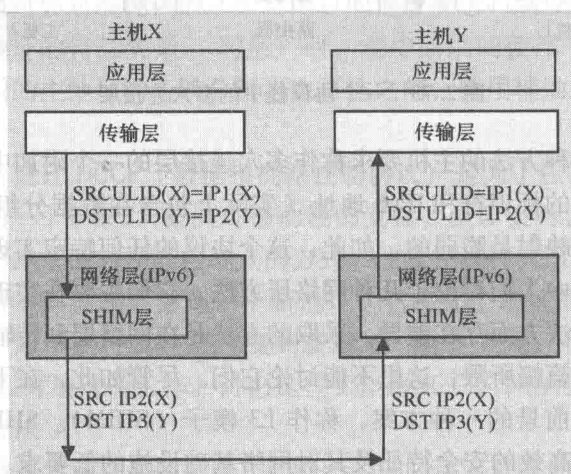


图 2.74 SHIM 方法图示

2）移动性和主机多穴连接

有趣的是在这里指出，保留通过主机移动性的已建立通信，类似于在多穴连接主机中通过中断保留已建立的通信。两种场景要求在通信期间具备使用的定位符的动态改变能力，同时维持上层协议（ULP）使用的端点标识符保持不变。因为 MIPv6（RFC 3775）已经提供这项要求的支持，保持移动期间已建立的通信，所以值得探索的是，它是否也可用来探索一个多穴连接环境中会话可存活能力。有关这一点的详细讨论超出了本章的范围。

2.5.3 站点多穴连接

如前面指出的，站点多穴连接来自于继承一个以上地址前缀的站点或有一个以

上外部接口的一个站点。在任何一种情形中，当处理这样的站点时，需要解决一些问题。它们可被枚举如下：

1）在站点连接带有多个前缀的情形中，如出于策略原因，需要解决如何使主机从一个特定前缀中选择哈希地址。

2）在站点连接带有多个接口的情形中，需要解决两个问题，即如何使一个工作/最佳附接点用于到达和/或外发流量以及如何处理这些附接中状态的改变。

3）在站点多穴连接带有多个接口的情形中，需要解决如何针对到达和/或外发流量在多个附接点间取得负载共享以及如何接收寻址到其接口已经宕机之前缀的分组。

4）在多个接口站点多穴连接的情形中，需要解决如何确保来自一个特定源前缀的分组通过那个前缀的附接点出口外发。

一些IPv6功能特征同样可被用来解决这些问题。例如，IPv6中的一个接口可同时有一个全局地址和一个站点本地地址（像IPv4“net10”地址）和基于交换的寻址，它将一个顶级汇聚标识符（TLA）指派到一组互相连接的ISP。

下面将讨论焦点放在参考文献中出现的站点多穴连接解决方案[76]上，注意这样的事实，即IPv6中的多穴连接还没有标准化。

1. IPv4中的站点多穴连接

对于数量日渐增长的公司而言，因特网连接能力占有战略地位。所以，许多公司的网络希望连接到至少两家ISP而到达因特网，基本出发点，不仅是在ISP的网络出现故障事件时增强它们的可靠性，而且在于增加它们的网络性能，如网络延迟。所以，最近站点多穴连接正取得重要地位。在如今的IPv4因特网中，至少60%的ERD以多穴方式连接到两个或多个ISP，且它们的数量正在增加。许多站点期望在IPv6中是多穴连接的，即使带有到全球移动通信系统（GSM）多个接口的端用户、通用移动通信系统（UMTS）或802.11网络也是如此。

IPv4中站点多穴连接的传统方法是使用BGP，向其每个ISP通告单个站点前缀。在采用BGP的站点多穴连接中，一个特定站点可使用提供商无关的（PI）或提供商汇聚的（PA）地址。下面在图示的帮助下将讨论这些方法。

（1）采用提供商无关地址的站点多穴连接

在这种情形中，一个多穴连接站点得到提供商无关的一个前缀地址，并使用BGP将之通告到其ISP。对于IPv4中的站点多穴连接，这是首选方式，原因是如果站点改变提供商，则站点不需要重新编址。同样，直到20世纪90年代中期，任何站点从区域因特网注册机构（RIR）得到一个相当大的PI地址空间，还是相对容易的。考虑图2.75中给出的场景。

这里，ERD 112是足够大的，可从RIR得到一个PI前缀。ERD 112使用BGP，将其前缀通告到它的两个提供商TRD 40和TRD 20。接下来，这两个TRD（除了它们自己的前缀外）将ERD 112的前缀通告到全球因特网。这为因特网的其他部

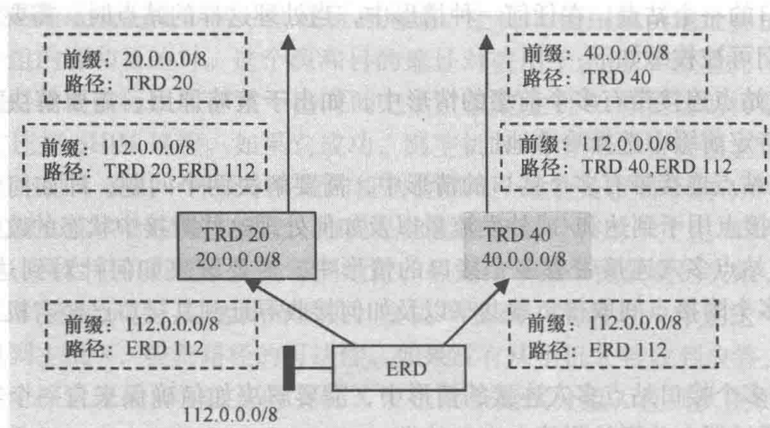


图 2.75　使用 PI 地址的站点多穴连接

分提供了回到多穴连接站点的多条路径。采用这种方式，一条附加的路由表项被引入到全球路由系统。采用广泛使用的站点多穴连接方法，这将导致扩展性问题。同样由于 IPv4 地址空间的快速耗尽，从 RIR 得到一个无关前缀是不容易的。所以，这些原因使这个规程不再是一个有吸引力的提案。

（2）使用提供商汇聚地址的站点多穴连接

在这种情形中，多穴连接站点使用提供商之一汇聚它的前缀地址[63]。通过考虑如图 2.76 所示的场景，这可良好地得到说明。

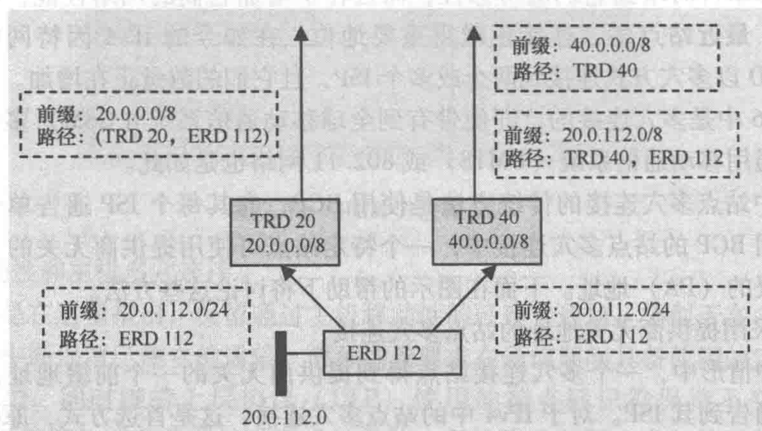


图 2.76　使用 PA 地址的站点多穴连接

这里，ERD 112 使用单个 PA 地址空间。这个地址空间是由主中转提供商 TRD 20 指派的。ERD 112 使用 BGP 将前缀 20.0.112.0/24 通告到两个提供商 TRD 20 和 TRD 40。注意在这种情形中，TRD 20 能够以其自己的 20.0.0.0/8 前缀汇聚 ERD 112 的前缀，TRD 20 仅将可汇聚的地址通告到全球因特网。同时，因为 TRD 40 不

能汇聚 ERD 112 的前缀，所以与其自己的前缀（40.0.0.0/8）一起，它通告 ERD 112 前缀（20.0.112.0/24）。

这项规程的缺陷是，除了在全球路由表中引入一条附加的路由表项外，无论何时站点改变主中转提供商，多穴连接的站点都需要进行重新编址。在过去数年，考虑到因特网的指数增加（这接下来增加了路由表的尺寸），这些站点多穴连接解决方案导致运营问题，并对网络性能造成一些负面影响。所以，对于下一代 IPv6 网络，要求基于路由汇聚的新解决方案。

2. IPv6 中的站点多穴连接

如前面提到的，在当前世界中，多数站点请求进行多穴连接，以便保护它们免受其提供商链路失效的影响，且 IPv6 被设计为支持这些设施工具，原因是对这种服务的需要正在增长。有时，一个站点采纳多穴连接，目的是将其流量分布在多个中转提供商间，以便就时延、抖动、分组丢失和带宽方面取得更好的网络性能。这意味着 IPv6 站点多穴连接的任何解决方案应该包括两个显著的特征，即全容错和流量工程能力。全容错意味着一个 IPv6 站点多穴连接解决方案必须提供故障事件间的传输层存活能力。在处理整个巨大的下一代网络时，为满足负载共享和网络性能，非常需要流量工程功能。既然说到这一点，在开发任何 IPv6 站点多穴连接解决方案时，将存在技术和非技术方面的约束。正在这方面寻找 IPv6 站点多穴连接解决方案的主要约束如下：首先，在这方面的任何解决方案都需要针对因特网中 BGP 路由表的大小；第二，任何多穴连接解决方案不应出于安全原因而排除一个过滤规程；第三，多穴连接解决方案需要是提供商无关（即多穴连接无关性）；另外，任何多穴连接解决方案都必须对主机、路由器和 DNS 具有有限的影响，从而可容易地部署和操作。

在上述考虑的语境中，在参考文献中针对 IPv6 中的站点多穴连接，正在讨论两种主要方法。依据提供容错、流量工程、路由汇聚和多穴连接无关性的基础机制，对这些多穴连接解决方案进行分配。

（1）路由方法

这组 IPv6 站点多穴连接解决方案依赖于路由系统提供容错和流量工程功能。这样的路由机制包括 BGP 的使用、BGP 路由通告的过滤或域间隧道的使用。属于这个组的 IPv6 站点多穴连接解决方案有采用 BGP 的 IPv6 站点多穴连接、采用路由汇聚的 IPv6 站点多穴连接和使用提供商之间协作的 IPv6 站点多穴连接。将在下面讨论这些规程。

1）采用 BGP 的 IPv6 站点多穴连接

在采用 BGP 的 IPv6 站点多穴连接中，该解决方案类似于传统 IPv4 站点多穴连接解决方案，其中一个多穴连接站点可采用 PI 或 PA 地址。这里一个站点使用 BGP 将其自己的前缀通告给每个提供商[63,67]。图 2.77 形象地说明了采用 PA 地址的一种 IPv6 站点多穴连接解决方案。容错和流量工程是由 BGP 和 IGP 的充分配置提供

的。但注意，如前面所解释的，这种解决方案导致扩展性问题，原因是每个多穴连接的站点在因特网中的所有路由器的 BGP 路由表中引入一个新的前缀。

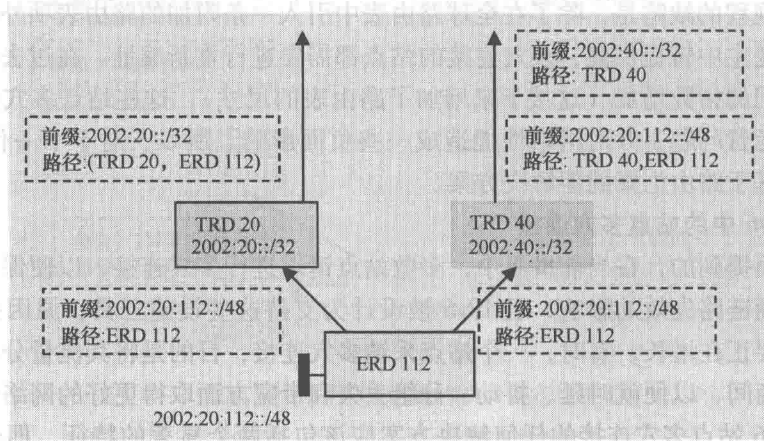


图 2.77　使用 PA 地址的 IPv6 站点多穴连接

2）采用路由汇聚的 IPv6 站点多穴连接

这种站点多穴连接解决方案依赖于提供商，他们协作过滤 BGP 路由以支持路由汇聚，同时提供某种容错。它使用现有协议和实现[63,67]。图 2.78 形象地说明了这个规程。

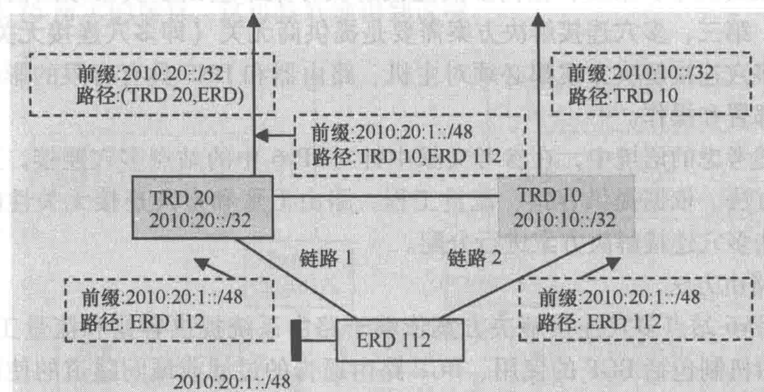


图 2.78　采用路由汇聚的 IPv6 站点多穴连接

这里，多穴连接站点 ERD 112 被连接到提供商 TRD 10 和 TRD 20。ERD 112 从 TRD 20 得到单个 PA 前缀 2010：20：1::/48，TRD 20 也是主 ISP。ERD 112 使用 BGP 将其前缀 2010：20：1::/48 通告到 TRD 10 和 TRD 20。为首选域间路由，TRD 10 仅传播 ERD 112 的前缀到 TRD 20。TRD 20 能够以其自己的前缀（2010：20::/32）汇聚 ERD 的前缀（2010：20：1::/48），并仅将汇聚地址通告到全球因特网。注意在这种情形中，它不将 2010：20：1::/48 传播到全球因特网。作为这

种情况的结果，来自因特网且目的地为多穴连接站点的流量，总是通过 TRD 20 路由。之后 TRD 20 将目的地为多穴连接站点的流量直接通过链路 1 转发，或依据一些路由策略通过 TRD 10 转发。ERD 112 可以不同方式发送它的外发流量，通过 TRD 10 或 TRD 20。如果链路 2 失效，则进入和外发流量将通过链路 1 流动。如果链路 1 失效，则进入流量将通过路径 TRD 20 TRD 10 ERD 112 到达 ERD 112，而外发流量以相反顺序流动。

这种解决方案的主要缺陷如下：首先，在失效的情况下，它不能在主 ISP 和连接到多穴连接站点的链路内提供容错；其次，如果在提供商 TRD 10 和 TRD 20 之间没有直接链路，则前缀 2010：20：1 : :/48 必须通过中间的中转提供商进行传播，而这种协作可能与他们的商务（规则）是冲突的。

3）在站点出口路由器处支持多穴连接的 IPv6 站点

这种站点多穴连接解决方案基于隧道和多个前缀的使用（RFC 3178）[64]。多穴连接站点依据每个提供商指派一个前缀。图 2. 79 形象地说明了这个规程。

这里，依据提供商 TRD 20 和 TRD 10，ERD 1120 已经分别被指派了前缀 2010：20：1 : :/48 和 2010：10：1 : :/48。这些前缀分别由站点出口路由器 RA 和 RB 通告到 ERD 1120 内部的每台主机。通过仅将这个提供商分配的前缀通告到给定的提供商，做到路由汇聚。所以，每个提供商能够实施路由汇聚。例如，ERD 1120 仅将前缀 2010：20：1 : :/48 通告给 TRD 20，TRD 仅将其自己的 IPv6 汇聚 2010：20 : :/32 通告到全球因特网。通过使用辅助线路（通常是 IP 隧道），在 RA 和 TRD 10 以及 RB 和 TRD 20 之间建立链路，提供冗余。RA 仅在辅助链路上将其前缀 2010：20：1 : :/48 通告到 TRD 10，而 RB 仅在辅助链路上将其前缀 2010：10：1 : :/48 通告到 TRD 20。

这个架构提供路由汇聚，并能够在链路失效间保留已建立的 TCP 连接。但这个规程的主要缺陷是，它不能处理任何上行 TRD 的失效，并强制每个 TRD 配置隧道。

(2) 中间设备的方法

通过在多穴站点间由中间设备提供的服务，这些方法提供站点多穴连接功能。属于这个类的解决方案包括采用 NAT 的站点多穴连接、站点多穴连接别名协议和站点多穴连接转换协议。但是，一般而言，认为这些方法不适合于 IPv6，原因是这些方法破坏了因特网的端到端原则，导致许多安全问题。所以，出于完备性考虑，仅讨论这些方法中的一种方法。要了解其他方法，请参见参考文献 [76，81]。

采用 NAT 的 IPv6 站点多穴连接

这种方法依赖于使用 NAT[60]，将分组定向转发到正常工作的提供商。一般而言，NAT 路由器被安装在网络边缘，且它非常了解哪个提供商在工作，哪个提供商出问题了。

在这个信息的基础上，它以属于可正常工作的提供商的一个 IP 地址替换外发

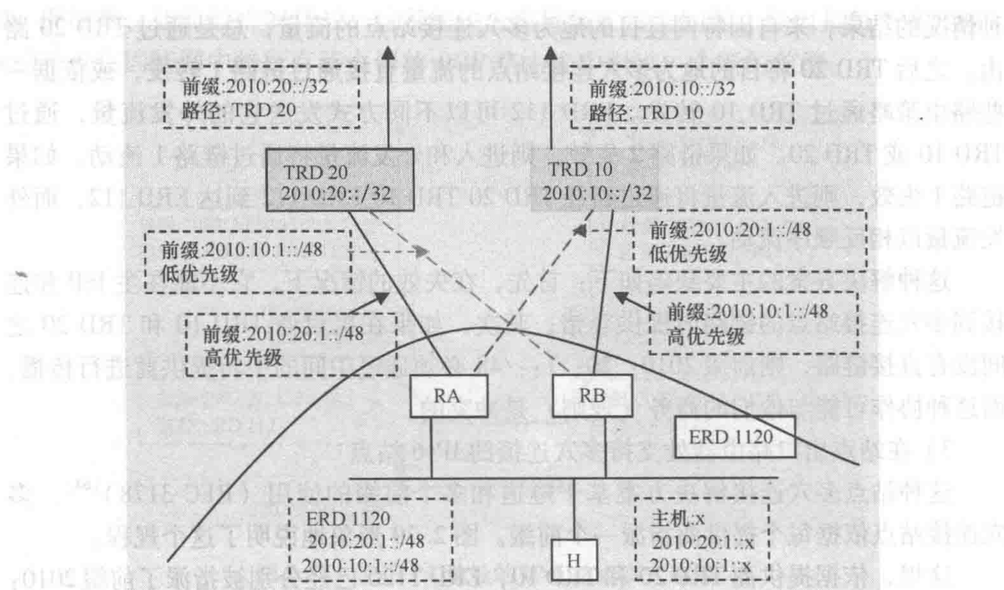


图 2.79　在站点出口路由器处带有多穴连接支持的站点

分组的 IP 地址。图 2.80 形象地说明了以 NAT 实现多穴连接的一个站点。该站点有两个 IPv6 前缀，每个前缀来自一个提供商。在站点内的一台主机可使用任意提供商的地址。这种站点多穴连接解决方案[77]支持路由汇聚，提供与提供商的完全独立性，且不需要 BGP，但它破坏了 IPv6 的端到端原则，导致重大的安全问题。

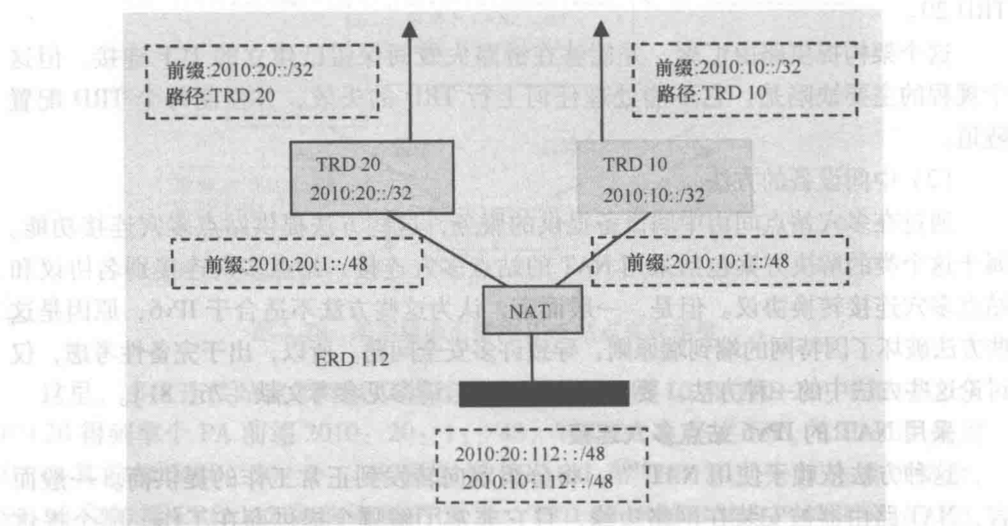


图 2.80　采用 NAT 的 IPv6 站点多穴连接

2.6 移动性

对未来因特网的移动性支持是至关重要的，原因是移动多媒体通信正快速地扩散开来，且移动计算已经成为如今的常态。节点或主机被设计为移动的和快速移动的（Move and Move Fast）。一些节点是偶然移动的，而其他节点则所有时间都在移动，并要求持续的网络连接能力。例如，在诸如保健、销售、保险和战场军力等业界中的移动用户，都正在寻找网络连接能力"总是在线"的网络连接优势，以便接收连续的商务上至关重要的数据。标准移动IP［移动IPv4（MIPv4）和MIPv6］指IP的移动性方面，使节点可移动到不同网络，同时维持高层连接能力（就像在GSM中无线话音通信中发生的情况，此时节点从一个网络移动到另一个网络，也称作无缝切换）。这不要与"便携性"混淆，后者使节点移动到世界所有地方的不同网络并保持可达的，但高层连接在每次节点重新定位时一定会被中断，原因是它必须在每个地点得到一个新的地址，并由该地址加以寻址。

因特网中的移动性是困难的，原因是在如今的因特网中路由和识别是相关的，这一点不像无线电信网络。一个IP节点是由其IP地址标识的，它由一个网络部分和一个主机或接口部分组成。在如今的因特网中路由是使用IP地址的网络前缀完成的。如果一个节点在保持相同地址的情况下移动到另一个网络，则分组将在节点IP地址网络部分的基础上被路由到其原网络，所以将永远不能到达节点新网络位置的该节点处。但是，如果一个节点移动到另一个网络，并基于新网络的网络前缀得到一个新地址，虽然它将接收到发送给它的所有后续分组，但它将丢失从其原网络中与对端已经建立的会话。所以，移动IP的目标是支持一个移动节点（MN）保持相同地址，而不管其到因特网的联系点，以便维持现有连接，同时在因特网中的任何新位置保持为可达的。虽然对移动性的需求是不可否认的，但为这样的需求提供必要支持的现实却给出几项挑战，包括：

1）IP地址的可用性。

2）由于折中处理方法（Workaround）（如NAT和链路层切换机制）导致当前网络技术的复杂性。

3）在全球范围上对"总是在线"和安全通信的需要。

4）在多址环境中访问网络的需要。

5）低访问速度以及竞争的和不兼容服务的负面效果。

本节的主要目标是理解移动IPv4（MIPv4）的基本工作原理以及在IPv6中是如何修改的，还有这样做的优势。也会了解移动IPv6（MIPv6）中的基本操作。在开始时，希望明确的是，没有在本章涵盖有关IPv6移动性的任何高级专题。欲了解高级专题，请读者参见参考文献［87，88］。

2.6.1 移动 IPv4

开始时指出，IPv4 没有内建移动性。为支持移动性，将 MIPv4 设计为基本 IPv4 协议的一个扩展。MIPv4[85,87]解决移动性问题，方法是在每个新的位置为一个 MN 指派一个临时地址（遵循蜂窝移动网络的步骤），维持该 MN 的原 IP 地址，采用 MN 原网络中的一台路由器（称作本地代理）创建并存储这两个地址间的一个绑定［遵循蜂窝移动网络的归属位置寄存器（HLR）的步骤］。一个 MN 在其原（本地）位置从 HA 得到一个 IP 地址，称作本地地址，并为维护端到端通信而保有这个地址，在每次它移动到一个新网络时，出于路由目的，它也从一个外地代理（FA）（在一个新位置处的一台路由器）处得到一个临时地址或转交地址（CoA），类似于 MN 从 VLR 得到临时移动站标识符（TMSI），无论何时它进入蜂窝移动网络中的一个新网络时。MN 发送一条更新，称作绑定更新（BU），包含到其 HA 的新 CoA，这允许 HA 为 MN 在其本地地址及其 CoA 之间创建一条绑定（类似于蜂窝移动网络中 HLR 的功能）。使用这个绑定缓存，HA 截获目的地为 MN 本地地址的任何分组，采用其 CoA 封装，并以隧道方式将之传输到在外地位置的 MN。由此，通过 HA 为 MN 维持的绑定，使用 MN 的 CoA，它在因特网中任何位置时都是可达的，且其移动对使用本地地址的高层应用是透明的，如图 2.81 所示。

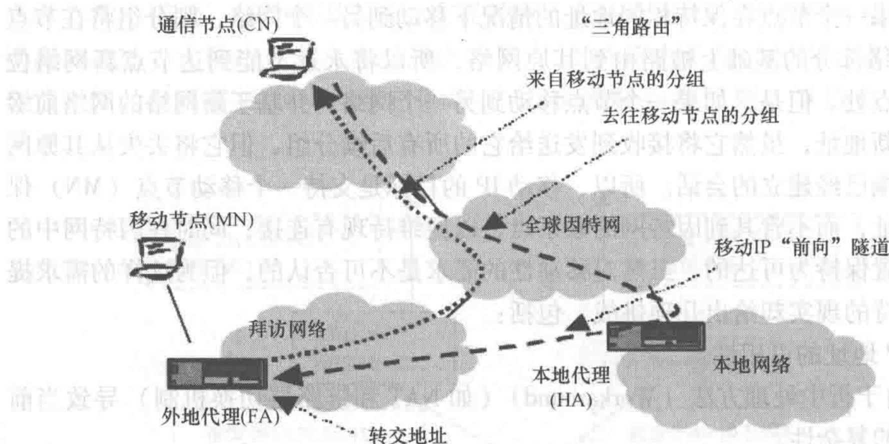


图 2.81 MIPv4

介绍 IPv4 中基本移动操作考虑两种情形，即一个对端［通信节点（CN）］希望与 MN 通信，以及 MN 希望与 CN 通信（见图 2.81）。注意，无论何时 CN 希望发送分组到 MN 时，如果 MN 离开本地，则在到达 MN 之前，分组必须传输到本地网络。这种低效路由被称作"三角路由"。在 MN 发送分组到 CN 的情形中，分组以层 2 技术发送到 FA。因为假定 CN 是有一个公开可路由地址的，所以 FA 就可能直接将分组转发到 CN，这就旁路了 HA。

注意，虽然该技术似乎是直接的和简单的，但它有几个固有问题。例如：

1）规程不能以一种良好方式支持私有寻址，原因是该解决方案要求在每个接口上有唯一的 IP 地址。

2）因特网路由器严格地过滤不是从一个拓扑上正确的子网发出的分组。在这样一种情形中，MN 的分组不能由 FA 转发到 CN。为解决这个问题，人们可能重新回到从 FA 到 HA 的反向隧道，而不是直接将分组发送到 CN。但这同样是一种低效路由，类似于三角路由。

在参考文献中人们提出了 MIPv4 中涉及的大量解决方案和规程细节，其讨论在这里有点偏题。欲了解附加阅读材料，感兴趣的读者可参见参考文献［82，83］。

2.6.2　移动 IPv6

注意，为带来无线因特网移动能力，在 IPv6 设计中，移动性是一项核心功能特征，而在 IPv4 设计中，它是一项事后补救措施。出于这个原因，MIPv6 有优于 MIPv4 的几项核心操作优势。在下面将研究这些优势。

1. 移动 IPv6 设计优势

这里列出 MIPv6 优于 MIPv4 的一些设计优势[86]。

1）较大型的地址空间：MIPv4 中的 FA 提供可为多个 MN 共享的一个 CoA。这意味着，FA 去除了向每个 MN 指派一个唯一的、共位 CoA 的需要。但是在 IPv6 中，地址的可用性不是一个问题，IPv6 支持多达 $2^{128}=3.4028237\times10^{38}$ 个可寻址节点（注意 10^{12} 等于 10000 亿）。同样，这个巨大的地址空间支持地址的非常简单的自动配置，并支持一个 MN 在任意外地链路上快速地和容易地获取一个共位 CoA。结果，FA 功能在 MIPv6 中是简单的，同样 CoA 的 FA 变形也是简单的。所以，在 MIPv6 中的唯一 CoA 类型是共位 CoA。所以，不需要像在 MIPv4 中一样在 MIPv6 中部署特殊路由器作为 FA。MIPv6 工作在任意地点，不需要来自本地路由器的任何特殊支持。

2）新路由首部（RH）：IPv4 松散源和记录路由选项的令人不期望的特征，是接收到包含这些选项的一个分组的各节点，当对该分组的原始源做出应答时，要反转选项。这为简单的拒绝服务攻击打开了大门。在 RFC 1883 中定义的 IPv6 RH，没有处理这个性质，即接收到包含一个 RH 的一条 IPv6 分组的节点，当对原始源做出应答时，不需要包括该 RH。其次，包含任何选项的 IPv4 分组必须由沿其路径的每台路由器进行检查，这使这种分组的转发相对低效。相对而言，IPv6 显性地将选项分为必须由每台路由器检查的那些选项和仅需由最终目的地检查的那些选项。由此，IPv6 RH 可完全由沿路的多数路由器忽略，这在多数这样的路由器中支持非常快速的决策。

3）认证首部：此外，IP 认证首部的实现对于 IPv6 节点是必需的。这也许为路由优化技术的大规模采用提供一种机制。MIPv6 路由优化（在没有 HA 干扰情况

下，CN和MN之间的直接通信）可安全地操作，即使没有预先安排的安全关联也如此。预计，路由优化可在所有移动节点CN间的全球规模上部署。

4）路由优化：MIPv4路由优化是协议的一个扩展，而不是基础RFC 3775[86]的组成部分。它要求预配置和静态安全关联；难以操作入口过滤（Ingress-Filtering）路由器（MIPv6路由优化是包括在这个协议中的一个基础部分）；它被集成可返回到动态安全路由优化的可路由能力；它可与入口过滤路由器高效地工作。

5）与链路层解耦：MIPv6与任何特定的链路层解耦，因为它使用IPv6邻居发现而不是IPv4 ARP。这也改进了该协议的鲁棒性。

6）移动扩展首部：为MIPv6信令消息（如BU、本地地址和CoA）和绑定请求扩展IPv6移动扩展首部。

7）动态HA地址发现：MIPv6中的动态HA地址发现机制将单条应答返回给MN。在IPv4中使用的定向广播从每个HA返回独立的应答。

8）目的地选项：扩展IPv6目的地选项，是为了包括一个MN的本地地址。

9）新的ICMPv6消息：新的ICMPv6消息用于HA发现请求和应答，以及前缀请求和通告。

2. 移动IPv6和移动IPv4：一种比较

MIPv6借鉴了MIPv4的许多概念和术语，如表2.6所示。例如，仍然有MN和HA，但没有FA。本地地址、本地链路、CoA和外地链路的概念大略与MIPv4中的相同。MIPv6利用隧道和源路由将分组交付到连接到一条外地链路的各MN，隧道是MIPv4中使用的唯一机制。MIPv6的高层功能与MIPv4的相同，并大略对应于MIPv4的三个组件：代理发现、注册和路由。

表2.6 MIPv6和MIPv4之间的比较

MIPv4概念	等价的MIPv6概念
MN、HA、本地链路、外地链路	相同
MN的本地地址	全局可路由本地地址和本地链路本地地址
FA	在外地链路上的一台“纯”（Plain）IPv6路由器（FA不再存在）
FA CoA 共位CoA	所有CoA都是共位CoA
通过代理发现、DHCP或手工的方式得到CoA	通过SLAAC、DHCPv6或手工的方式得到所有CoA
代理发现	路由器发现
采用HA的认证注册	HA和其他通信方的认证通知
通过隧道，路由到各MN	通过隧道和源路由，路由到各MN
通过协议规范的路由优化	对路由优化的集成支持

3. 移动 IPv6 操作

MIPv6 被设计为管理无线和 IPv6 网络之间一台 MN 的活动。当一台 MN 停留在其本地网络中时，它像另一个 IPv6 节点一样与其通信方通信。当一台 MN 移动到另一个子网中的一个新的附接点时，其本地地址不再有效，由其通信方发送的分组将继续到达其本地网络。因此，它需要在访问子网中得到一个新的有效地址，称作 CoA，并将其注册到它的 HA 和通信方。在一个 MN 的本地地址和当前 CoA 之间做出的关联，被称作一个绑定。由此，本地地址总是识别一台 MN 的通信，而 CoA 定位该 MN。图 2.82 ~ 图 2.84 给了出三个场景，说明了一台 CN、一台 HA 和一台 MN 之间的交互通信。一台 MN 使用 IPv6 版本的路由器发现[86,88]确定其当前位置。参照这三个场景和两种情形将解释 MIPv6。

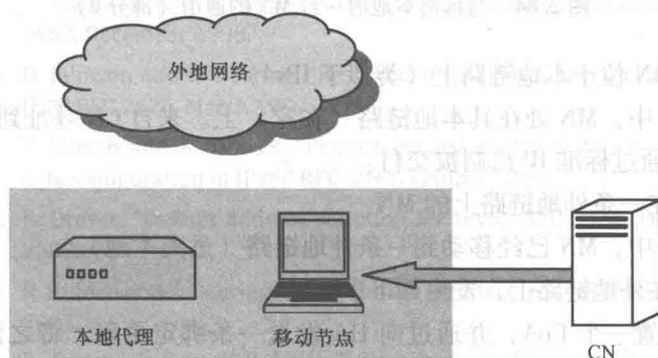


图 2.82　与在本地网络中的一台 MN 的通信

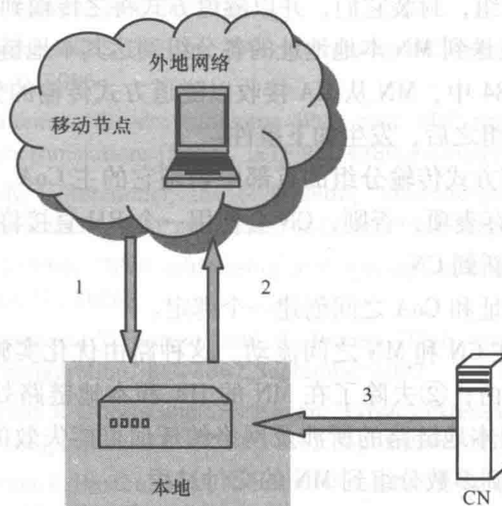


图 2.83　与远离本地的一台 MN 的通信（部分 1）

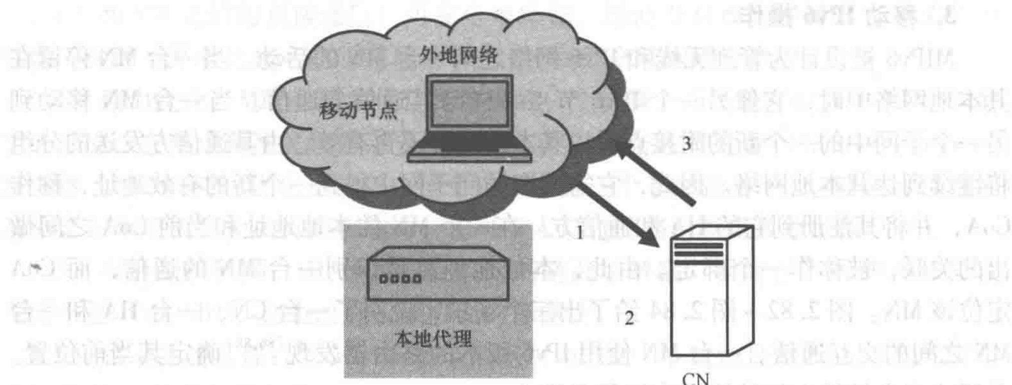


图 2.84　与远离本地的一台 MN 的通信（部分 2）

情形 1：MN 位于本地链路上（类似于 IPv4）

在图 2.82 中，MN 处在其本地链路（在家）上。来自 CN 寻址到 MN 的本地地址的各分组，通过标准 IP 机制被交付。

情形 2：在一条外地链路上的 MN

在图 2.83 中，MN 已经移动到一条外地链路（远离本地）。

部分 1：在外地链路上，发生如下事件：

1）MN 配置一个 CoA，并通过向 HA 发送一条绑定更新，将之注册到其 HA。这个新地址是 MN 的主 CoA。

2）通过将一条绑定确认返回给 MN，HA 确认绑定更新。

3）HA 截获各分组，封装它们，并以隧道方式将之传输到 MN 所注册的 CoA。

4）由一个 CN 发送到 MN 本地地址的各分组到达其本地链路。

部分 2：在图 2.84 中，MN 从 HA 接收以隧道方式传输的分组。在 MN 接收到以隧道方式传输的分组之后，发生如下事件：

1）MN 在以隧道方式传输分组的首部中识别它的主 CoA。MN 假定原发送 CN 没有该 MN 的绑定缓存表项，否则，CN 会使用一个 RH 直接将分组发送到 MN。之后它发送一条绑定更新到 CN。

2）CN 在本地地址和 CoA 之间创建一个绑定。

3）各分组直接在 CN 和 MN 之间流动。这种路由优化实施如下操作：①去除了普遍所谓的三角路由；②去除了在 MN 的 HA 和本地链路处的拥塞；③降低了 HA、本地链路或进出本地链路的所涉及网络的任何可能失效的影响，原因是这些节点和链路没有牵涉到多数分组到 MN 的交付过程。

当 MN 离开本地时，它总是发送一个本地地址选项，通知其本地地址的接收方。采用那种方式，接收方可正确地识别分组所属的连接。当 MN 返回到其本地链路时，MN 向 HA 和 CN 发送一条绑定更新，清除绑定关系。

参考文献

1. J. Bound, B. Carpenter, D. Harrington, J. Houldsworth, and A. Lloyd, "OSI NSAPs and IPv6," RFC 1888, August 1996.
2. R. Hinden and S. Deering, "IPv6 address architecture," IETF RFC 2373, July 1998.
3. R. Hinden, M. O' Dell, and S. Deering, "An IPv6 aggregatable global unicast address format," RFC 2374, July 1998.
4. S. Deering and R. Hinden, "Internet protocol version 6 (IPv6) specification," RFC 2460, December 1998.
5. S. Thomson and T. Narten, "IPv6 stateless autoconfiguration," RFC 2462, December 1998.
6. D. Johnson and S. Deering, "Reserved IPv6 subnet anycast address," IETF RFC 2526, March 1999.
7. T. Narten and R. Draves, "Privacy extensions for stateless address autoconfiguration in IPv6," RFC 3041, January 2001.
8. R. Draves, "Default address selection for IPv6," RFC 3484, February 2003.
9. R. Hinden and S. Deering, "IPv6 address architectures," IETF RFC 3513, 2003.
10. R. Droms, J. Bound, B. Volz, T. Lemon, C. Perkins, and M. Carney, "Dynamic host configuration protocol for IPv6," RFC 3315, July 2003.
11. R. Hinden and B. Haberman, "Unique local IPv6 unicast addresses," IETF RFC 4193, October 2005.
12. R. Hinden and S. Deering, "IPv6 addressing architecture," RFC 4291, February 2006.
13. C. Huitema, "Teredo: tunneling IPv6 over UDP through network address translations (NATs)," IETF RFC 4380, February 2006.
14. E. Gray, J. Rutemiller, and G. Swallow, "Internet code point (ICP) assignments for NSAP addresses," IETF RFC 4548, May 2006.
15. Tim Rooney, "IPv6 addressing and management challenges," BT Diamond IP, 2006.
16. S. Thomson, T. Narten, and T. Jinmei, "IPv6 stateless address autoconfiguration," RFC 4862, September 2007.
17. T. Narten, R. Draves, and S. Krishnan, "Privacy extension for stateless autoconfiguration in IPv6," RFC 4941, September 2007.
18. T. Narten, G. Huston, and L. Roberts, "IPv6 address assignments to end site," RFC 6177, March 2011.
19. Microsoft TechNet, "IPv6 addressing," http://technet.microsoft.com/enus/library/cc775951(WS.10).aspx.

20. IPv6.com, "IPv6 addressing," http://ipv6.com/articles/general/IPv6-Addressing.htm.
21. Mohamed G. Gouda, *Elements of Network Protocol Design*, John Wiley and Sons, 2004.
22. S. Deering and R. Hinden, "Internet protocol version 6 (IPv6) specification," RFC 2460, December 1998.
23. B. Gandalf, D. Carr Newbridge, and W. Simpson Daydreamer, "PPP Gandalf FZA compression control," RFC 1993, August 1996.
24. E. Nordmark, "Stateless IP/ICMP translation algorithm (SIIT)," IETF RFC 2765, February 2000.
25. G. Tsirtsis and P. Srisuresh, "Network address translation—protocol translation (NAT-PT)," IETF RFC 2766, February 2000.
26. K. Tsuchiya, H. Higuchi, and Y. Atarashi, "Dual stack hosts using bump-in-the stack technique (BIS)," IETF RFC 2767, February 2000.
27. R. Gilligan and E. Nordmark, "Transition mechanism for IPv6 hosts and routers," IETF RFC 2893, August 2000.
28. B. Carpenter and K. Moore, "Connection of IPv6 domains via IPv4 clouds," IETF RFC 3056, February 2001.
29. S. Lee, M-K Shin, Y-J Kim, E. Nordmark, and A. Durand, "Dual stack hosts using bumps-in-the-API (BIA)," RFC 3338, October 2002.
30. Microsoft Corporation, "Introduction to IPv6," September 2003.
31. Microsoft Corporation, "IPv6 transition technologies," 2003.
32. E. Nordmark and R. Gilligan, "Basic transition mechanism for IPv6 hosts and routers," IETF RFC 4213, October 2005.
33. F. Templin, T. Gleeson, M. Talwar, and D. Thler, "Intra-site automatic tunneling addressing protocol (ISATAP)," IETF RFC 4214, October 2005.
34. Martin Dunmore, "An IPv6 deployment guide," 6NET Consortium, September 2005.
35. R. Hinden and S. Deering, "IPv6 addressing architecture," RFC 4291, February 2006.
36. C. Huitema, "Teredo: Tunneling IPv6 over UDP through network address translations (NATs)," IETF RFC 4380, February 2006.
37. T. Chown, "Use of VLANs for IPv4-IPv6 co-existence in enterprise networks," RFC 4554, June 2006.
38. Tim Rooney, "IPv4-to-IPv6 transition technologies," BT Diamond IP, 2007.
39. D. Shalini Punithavathani and K. Sankaranarayanan, "IPv4/IPv6 transition mechanisms," *European Journal of Scientific Research*, ISSN 1450-216X, 34(1), 110–124, 2009.
40. C. Hendrik, "Routing information protocol," RFC 1058, June 1988.

41. D. Estrin, Y. Rekhter, and S. Hotz, "A unified approach to inter domain routing," RFC 1322, May 1992.
42. Y. Rekhter and T. Li, "A border gateway protocol 4 (BGP-4)," RFC 1771, March 1995.
43. Y. Rekhter and P. Traina, "Inter-domain routing protocol version 2," June 1996.
44. G. Malkin, "RIP version 2," RFC 2453, November 1998.
45. T. Narten, K. Nordmark, and W. Simpson "Neighbor discovery for IP version 6 (IPv6)," RFC 2461, December 1998.
46. G. Malkin and R. Minnear, "RIPng for IPv6," IETF RFC 2080, January 1999.
47. P. Marques and F. DuPont, "Use of BGP-4 multiprotocol extension for IPv6 inter-domain routing," RFC 2545, March 1999.
48. R. Coulton, D. Fergusson, and J. Moy, "OSPF for IPv6," RFC 2740, June 1999.
49. T. Bates, Y. Rekhter, R. Chandra, and D. Katz, "Multiprotocol extension for BGP-4," RFC 2858, June 2000.
50. G. Huston, "Commentary on inter-domain routing in the Internet," RFC 3221, 2001.
51. R. Chandra and J. Scudder, "Capabilities advertisements with BGP4," RFC 3392, November 2002.
52. Douglas. E. Comer, *Computer Networks and Internet with Internet Applications* (4th edition), 2004.
53. "Routing with BGP4," Alcatel white paper, 2005.
54. Y. Rekhter, T. Li, and S. Hares, "A border gateway protocol 4 (BGP-4)," RFC 4271, January 2006.
55. J. Abley and P. Savola, "Depreciation of type 0 routing headers in IPv6," RFC 5095, December 2007.
56. T. Bates, R. Chandra, D. Katz, and Y. Rekhter, "Multiprotocol extensions for BGP-4," RFC 4760, January 2007.
57. R. Coltum, D. Fergusson, J. Moy, and A. Lindem "OSPF for IPv6," RFC 5340, July 2008.
58. S. Krishnan, "Handling of overlapping IPv6 fragments," RFC 5722, December 2009.
59. J. Arkko and S. Braner, "IANA allocation guidelines for IPv6 routing header," RFC 5871, May 2010.
60. P. Akkiraju and Y. Rekter, "A multihoming solution using NATs," Internet draft, 1998.
61. P. Fergusson and D. Senie, "Network ingress filtering: defeating denial of service attacks which employ IP source address spoofing," BCP 38, IETF RFC 2827, 2000.

62. R. Stewart, Q. Xie, K. Morneault, C. Sharp, H. Schwarzbauer, T. Taylor, I. Rytina, M. Kalla, L. Zhang, and V. Paxson, "Stream control transmission protocol," IETF RFC 2960, 2000.
63. J. Jieyun, "IPv6 multihoming with route aggregation," IETF Internet Draft, August 2000.
64. J. Hagino and H. Sydner "IPv6 multihoming support at site exit routers," IETF RFC 3178, October 2001.
65. G. Huston, "Commentary on the inter-domain routing in the Internet," RFC 3221, 2001.
66. G. Huston "Analyzing the Internet BGP routing table Internet protocol," Journal, 2001.
67. K. Lindqvist, "Multihoming in IPv6 multiple announcements of longer prefixes," IETF Internet draft, December 2002.
68. J. Abley, B. Black, and V. Gill, "Goals for IPv6 site-multihoming architectures," RFC 3582, August 2003.
69. A. Matsumoto, M. Kozuka, and K. Fuzikawa, "TCP multihoming options," Internet draft, October 2003.
70. C. Huitema, R. Draves, and M. Bagnulo, "Host-centric IPv6 multihoming," Internet Draft, February 2004.
71. R. Atkinson and S. Floyd, "IAB concerns and recommendations regarding Internet research and evolution," IETF RFC 3869, August 2004.
72. E. Kohler, M. Handley, and S. Floyd, "Datagram congestion control protocol (DCCP)," IETF Internet draft, November 2004.
73. F. Baker and P. Savola, "Ingress filtering for multihomed networks," IETF BCP 84, 2004.
74. M. Bagnulo, "Hash based addresses (HBA)," Internet draft, December 2004.
75. E. Nordmark and M. Bagnulo, "Multihoming L3 shim approach," Internet draft, January 2005.
76. G. Huston, "Architectural approaches to multihoming for IPv6," RFC 4177, September 2005.
77. M. Bagnulo, A. Garcia Martinez, and A. Azcorra, "Efficient security for IPv6 multihoming," *ACM Computer Communication Review*, 35(2), 61–68, 2005.
78. M. Bagnulo, A. Garcia Martinez, A. Azcorra, and C. de Launoise, "An incremental approach to IPv6 multihoming," *Computer Communications*, 2005.
79. I. Van Beijnum, "Shim6 reachability detection," IETF Internet draft, 2005.

80. M. Bagnulo, A. Garcia Martinez, J. Rodriguez, and A. Azcorra, "End site routing support for IPv6 multihoming," *Computer Communications*, **29**, 893–899, 2006.

81. C. de Launoise and M. Bagnulo, "The path towards IPv6 multihoming," Survey article in the Internet, 2007.

82. Charles E. Perkins, "Mobile IP," *IEEE Communications Magazine*, May 1997.

83. James D. Solomon, *Mobile IP: The Internet Unplugged*, PTR Prentice Hall.

84. N. Montavont and T. Noël, "Handover management for mobile nodes in IPv6 networks," *IEEE Communications Magazine*, August 2002.

85. C. Perkins, "Mobility support for IPv4," RFC 3344, August 2002.

86. D. Johnson, C. Perkins, and J. Arkko, "Mobility support for IPv6," RFC 3775, June 2004.

87. C. Perkins, "IP mobility support for IPv4," RFC 5944, November 2010.

88. C. Perkins, "IP mobility support for IPv6," RFC 6275, July 2011.

89. J. Finney, S. Schmidt, and A. Scott, "Mobile 4-in-6: a novel IPv4/IPv6 transitioning mechanism for mobile hosts," *IEEE Communications Society/WCNC*, 2005.

第3章　因特网云内部的路由

Dattaram Miruke

3.1　网络、因特网和层

如今到处都有网络。我们有社交网络、电视网络、无线电网络等。当然也有计算机网络以及最大的和所有网络的“老祖宗”——因特网。决定一个网络存在的主要原因之一是，通过网络相互连接的各主机之间的通信。一个网络的基础特点之一是在网络的各成员之间传递消息。在一个人群中，当沟通或传递一条消息到您周边的某个人时，您可与他或她以个人方式交谈（单播），或您可大喊，从而在可听距离内的每个人都得到该消息（广播、任意播、组播等，取决于所有对您的呼喊感兴趣的那些人），包括那个人。当您需要与在您的话音所及范围外的某个人通信时，您可使用中间设备，如一部电话、一台蜂窝电话或因特网，即使用某种中介将消息传递给另一个人。就网络的这个通用特点而言，计算机网络也不例外。计算机通信，或更准确地说是驻留在计算机或计算机集群上各应用进行通信。基本上而言，计算机网络使用通信的相同隐喻说法。在相互邻域内即连接到相同子网的各计算机，通过直接发送消息或分组到接收方或各接收方进行通信，并多次使用呼喊的隐喻法，即广播一条消息到子网上的所有计算机。但是，当各计算机不得不与不在邻域中的计算机通信时，它们不得不使用中介或中间计算机。这些计算机以各种名称指代，取决于它们在遍历驻留在不同子网中各计算机之间的通信方面所扮演的角色，或更准确地说它们的介入程度，被称为网桥、集线器、交换机和路由器。基本上来说，人类不擅长处理复杂性，特别是由于多层次细节混杂在一起导致的复杂性。例如，一辆汽车的司机不必一直思考每次他或她打轮或换档时发生在发动机中的细节方面的相互作用。基本上而言，他或她由汽车设计者提供了可操作的一个界面，且他或她有假定一辆汽车如何工作的一个模型，即当他或她挂入一个较高档位时，汽车加速，当他或她向右打轮时，汽车右转等。依据定义，它是一个模型，可表示或没有表示机器工作的实际方式。类似地，有在周边的世界或环境中各种事物运作的各种模型。这些模型意味着为简化事物，帮助理解和预测周边的世界。计算机和计算机网络是由人类社会形成的最复杂物件或机制之一。为在一个计算机网络实现的不同层次处理这种复杂性，已经形成协助我们的一个模型。记住，不是所设计的所有网络都遵循这个模型，但这个模型更准确地说是帮助我们理解和管理设计、构建和管理这些网络中所涉及的复杂性。一个模型仅是一个模型，可能没有具

有与实际现实的相似性，这就像由甚至一幅非常详细的地图所表示的实际风景，将不会与真实风景一样。

为这个目的开发的模型称作开放系统互连（OSI）（Arick Chapin）。也有基于因特网协议（IP）网络的另一个模型。将讨论这两个模型，并对之进行比较，看看上述设备（如交换机和路由器）在计算机网络中扮演何种角色。OSI 为实现网络协议提供了一个框架。一个协议是在一组通信实体间以某种方式交换消息的一个形式系统和机制。一个路由协议将路由器之间路由信息的正在进行的交换进行形式化。

当然，在人类网络中，有管理社会相互作用的协议。即使当拜访某个人或写信给某个人时，使用一个称谓，且在信件上使用一个地址，以确保信件被发出或路由到正确的位置。类似地，为在计算机间通信，它们需要自己的协议集。它们需要寻址方案，寻址各消息或分组（用来在一个网络上计算机间通信）。

多数模型或框架是真实世界系统的抽象。所以 OSI 模型也是这样的。OSI 模型有 7 层，每层代表网络的一个不同方面。这种分层方法带来的结果之一是，不同层可改变低层技术，并为上层和下层提供相同服务，由此将高层与低层中的改变隔离开了，如图 3.1 所示。其中，FTP 表示文件传输协议；API 表示应用编程接口；MAC 表示媒介访问控制；LLC 表示逻辑链路控制；ATM 表示异步传递模式；TCP/IP 表示传输控制协议/因特网协议；UDP 表示用户数据报协议；CSMA 表示载波侦听多路访问；HTTP 表示超文本传输协议；DNS 表示域名服务；SONET 表示同步光网



图 3.1　OSI 和因特网参考模型

络；SCTP 表示流控制传输协议。

虽然 OSI 参考模型被用来对联网的系统进行建模和教学原因，但非常常见的一个特殊模型称作 TCP/IP 或 DOD 模型，是用于网络协议和应用实际实现的一个模型。如传说的，在一个时间点，假定 OSI 网络栈会使世界变成一个更佳的场所，且所有事物都遵循 OSI 参考模型进行实现。但是，正如在真实世界中通常的情况一样，对任何解决方案的最终测试是它是否正常工作以及它有多快地成为可用的。因为直到 OSI 模型被确信的时点为止，已经存在大量专用网络解决方案，如 IBM SNA、DEC Net、Burroughs BNA（Chapin）等，而事实是，由各研究和学术机构提出并实现的基于 TCP/IP 的网络模型，开始站住脚，并扩展其范围。所以当开发 OSI 模型和人们尝试实际实现基于 OSI 模型的协议时，发现在实际实现中它是非常慢的，由此 TCP/IP 网络的实践能力赢得天下，且 TCP/IP 成为网络世界的专用标准。

在没有多少困难的情况下，能够将 TCP/IP 或 DOD 模型映射到 OSI 模型（Chapin），如图 3.2 所示。依据实现，多数应用倾向于模糊了应用、表示和会话层之间的边界。图示的中间列中的图表明 IP 或网络层作为低三层的黏合或绑定层，在物理层有众多种技术，所以到如今，对上层（包括应用[3]）有一个统一的 API 和前端。

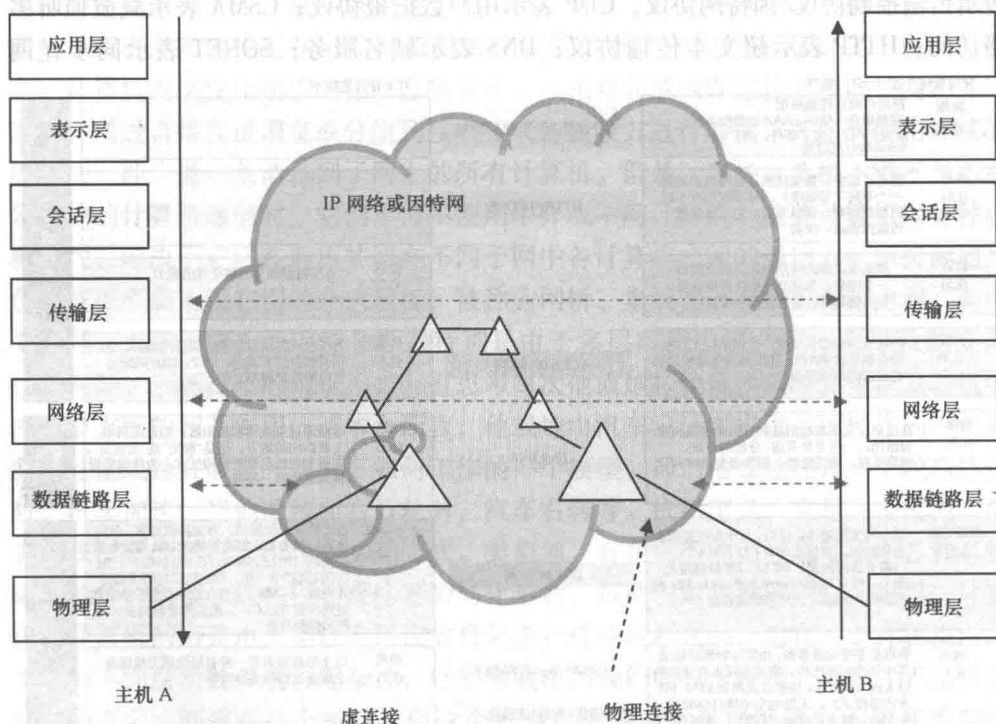


图 3.2 因特网云和路由器

3.1.1　层相互作用

在不讨论大量技术细节情况下，描述在真实生活中这些层是如何工作的一个通用例子。假定在用的协议栈是 TCP/IP，且用户将使用一个 FTP 客户端程序。为从一台 FTP 服务器得到文件或发送文件到该服务器，本质上将发生如图 3.3 所示过程。

3.1.2　因特网基础设施（因特网云内部是什么）

直到目前，您一定在想，因特网云内的节点来自哪里。如果在因特网的孩童时代您问了这个问题，那么我们会向您指出，大量节点实施任何两个点或主机（由因特网云连接的）之间多数分组的转发工作。但是，如今这涉及多个层的因特网服务提供商（ISP）。早期由政府资助的研究机构和大学所提供的功能，现在大部分由商业 ISP 所接管。基本上来说，因特网云是相互连接的各 ISP 的一项拼凑物，目的是转发流量或接收流量（Danny McPherson，1997）。

被问的一个非常频繁的问题是，因特网的“中心”在哪儿？好的，它的中心无处不在。因特网更像一组蜘蛛网（没有使用双关语）扩展超出一个房间，有点像覆盖整个区域，并沿它们的边连接起来，而没有任何中心点。基本上来说，因特网云是通过各种中转和对等安排（Peering Arrangements）相互连接的一项拼凑物，为的是发送和接收流量（Chapin）。

留给用户端的位置在哪儿呢？正常而言，用户端是这个层次结构中的端点，通过各种类型的物理连接而连接到一个 ISP，如拨号、数字用户线（DSL）、T-1 线路、宽带无线（Wi-Fi [802.11]）、微波接入的全球互操作能力 [WiMAX][802.16]、通用分组无线服务（GPRS 等）、线缆提供商、综合业务数字网（ISDN 等）（见图 3.4）。正常情况下，各 ISP 自己是通过使用 DSL、ATM、同步数字体系（SDH）（高速 DSL）或 10GB（或近期未来的 100GB）以太网连接进行连接的。各 ISP 有不同大小的规模，为了共享流量路由，他们与其他 ISP 有各种类型的布局安排（Arrangement）。

正常情况下，用户端与他或她从之得到因特网连接的 ISP 有一项支付约定。但是，各 ISP 自己必须为上游因特网访问付费，上游 ISP 有带有较大流量承载能力的一个较大型网络。直到到达少量层 1 承载商之前，重复这样的约定。在一些情形中，各 ISP 可能有一个以上的存在点，所以有到多个上游 ISP 的多条上游连接。

各 ISP 可使用多种类型的支付或优惠约定，如对等连接，其中多个 ISP 可互联。这些被称作对等点或因特网交换（IX）点（Brian Kahin），支持路由进行交换和流量（上行的和下行的）进行共享。正常情况下，层 1 ISP 仅有其他 ISP 作为客户，没有用户端客户。正常情况下对于因特网的现象，这是“理论联系实际（轮胎和马路接触）”（Rubber Meets the Road）的地方，即发生实际路由和流量共享。

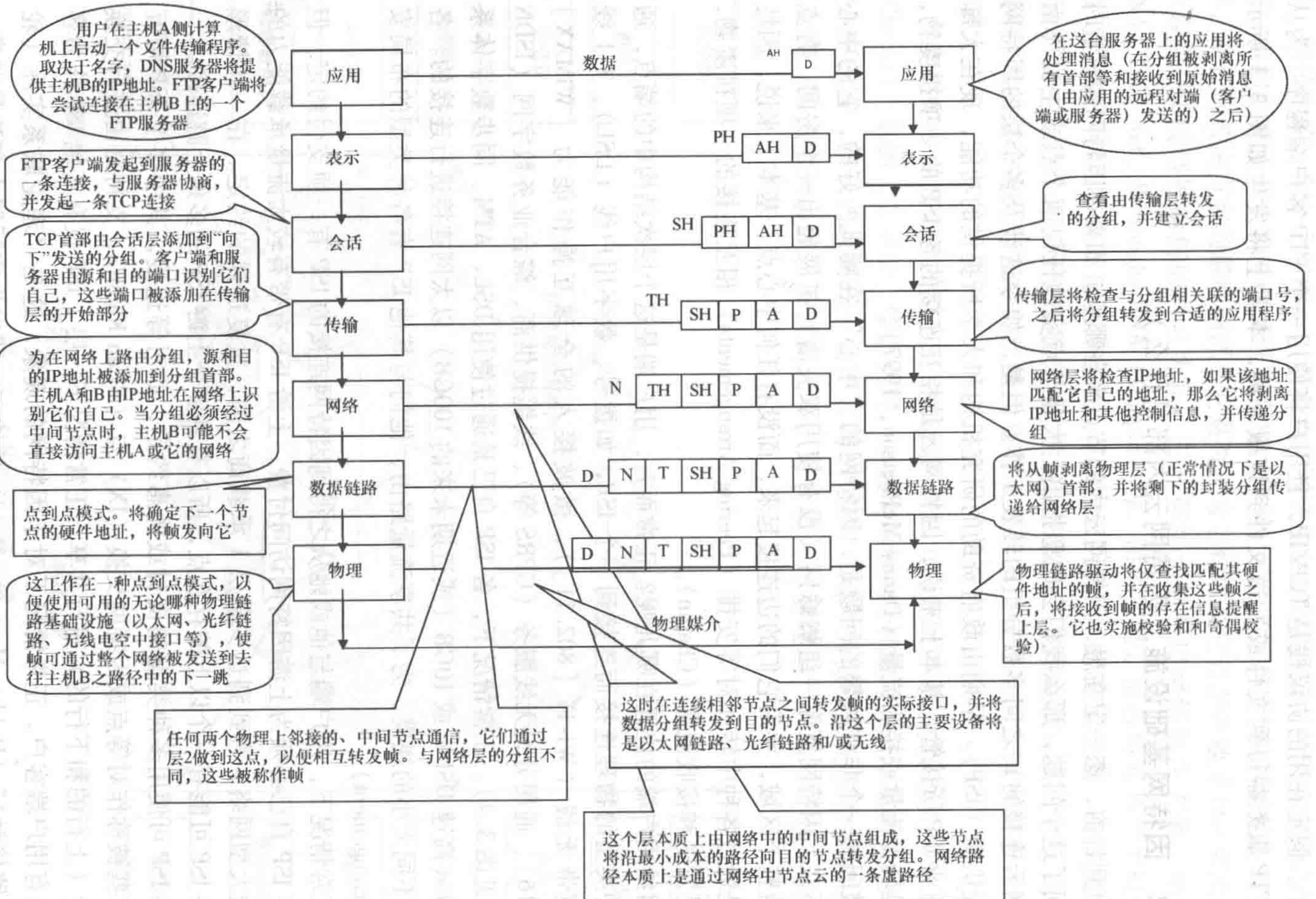


图 3.3　OSI 层的工作原理

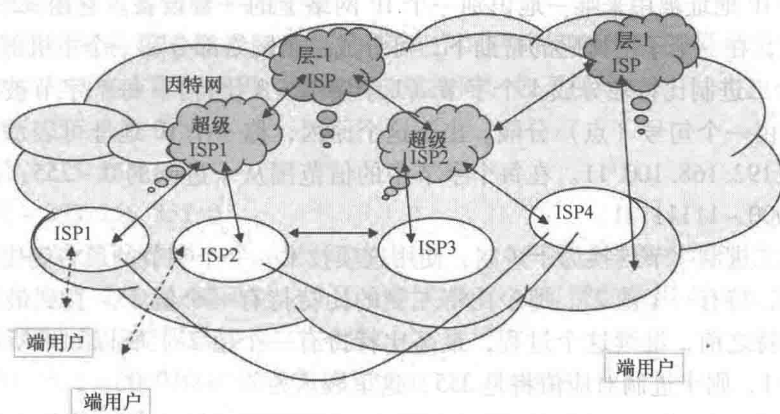


图 3.4　各 AS（自治系统，即各 ISP）互联并“形成因特网”

这是在开发和部署各种路由协议中所有专家知识发挥作用的地方。一个 IX 点是到因特网的一个访问点，并放有设备，如交换机、路由器和呼叫汇聚硬件（数字的和模拟的）。IX 点也被称作“共位中心”（黑色）。

对等连接更像大小相当的两个 ISP 之间流量的等量交换或交易，也被称作“无清算的对等连接”。其他约定可能包括中转（支付）给另一个网络，为因特网访问付费或作为一名客户，其中 ISP 网络因为因特网访问而付费。对等连接也可以是公开或私有对等连接的方式出现的。

如将在后面各节中看到的，边界网关协议（BGP）（Danny McPherson，1997）是用于各 ISP 之间路由流量的主要路由协议之一。所有这些“对等连接约定”的主要目的是提供全球可达性或端到端可达性。

3.2　网络和路由

本质上来说，一个网络是由多台设备之间的一个或多个边缘连接的一组设备。在 IP 网络的情形中，需要中间设备，这些设备将从一台主机将流量转发到一台或多台其他设备。流量可在前面看到的任意层处进行转发。在 IP 网络的情形中，用来转发流量的主要各层是层 1、层 2 和层 3，即物理层、数据链路层和网络层。本节将简短地描述用于在各主机间转发流量目的的一些设备。

3.2.1　IP 寻址

虽然预计本章的读者对 IP 寻址方案有一定程度的理解，但这里包括 IP 寻址和可变长度子网掩码（VLSM）的一个简短讨论，原因是当这个概念用在各种路由协议中的描述时，重要的是要理解这个概念。

一个IP地址被用来唯一地识别一个IP网络上的一台设备。它由32个二进制比特组成，在一个子网掩码的帮助下，可分成一个网络部分和一个主机部分。

32个二进制比特被分成4个字节（1个字节=8比特）。每个字节被转换为十进制，并由一个句号（点）分隔。出于这个原因，称一个IP地址可表示为点分十进制，如192.168.100.11。在每个字节中的值范围从十进制的0~255，或二进制的00000000~11111111。

为将二进制字节转换为十进制，使用这项技术。一个字节的最右侧比特或最低有效比特，持有一个值2^0。那个比特左侧的比特持有一个值2^1。直到最左侧比特或最高比特之前，继续这个过程，最高比特持有一个值2^7。所以如果所有二进制比特都为1，则十进制对应值将是255，这里表示为

1 1 1 1 1 1 1 1

128 64 32 16 8 4 2 1 (128+64+32+16+8+4+2+1=255)

这里有一个范例字节转换，此时不是所有比特都设置为1。

0 1 0 0 0 0 1 1

0 64 0 0 0 0 2 1 (0+64+0+0+0+0+2+1=67)

这个范例给出以二进制和十进制表示的一个IP地址。

10.0.31.23（十进制）

0001010.00000000.00011111.00010111（二进制）

存在5个不同类的网络，即A~E。A类~C类被大量使用，但D类和E类被保留。但是，在引入无类域间路由（CIDR）之后，这些类不再用于IP寻址。

给定一个IP地址，其类别可由高3比特确定。图3.5给出了高3比特的意义和落入每个类别的地址范围。



图3.5 IP编址方案

网络掩码和可变长度子网掩码

一个网络掩码帮助您了解地址的哪部分识别网络和地址的哪部分识别节点。A、B 和 C 类网络有默认掩码，也称作自然掩码，如下所示：

A 类：255.0.0.0

B 类：255.255.0.0

C 类：255.255.255.0

在没有划分子网的一个 A 类网络上的一个 IP 地址，将有类似于 8.20.15.1 255.0.0.0 的一个地址/掩码对。为了弄明白掩码如何帮助您识别地址的网络和节点部分，将地址和掩码转换为二进制数。

8.20.15.1 = 00001000.00010100.00001111.00000001

255.0.0.0 = 11111111.00000000.00000000.00000000

一旦将地址和掩码表示为二进制，那么识别网络和主机 ID 就比较容易了。将相应掩码比特设置为 1 的任何地址比特，代表网络 ID。将相应掩码比特设置为 0 的任何地址比特，代表主机 ID。

8.20.15.1 = 00001000.00010100.00001111.00000001

255.0.0.0 = 11111111.00000000.00000000.00000000

网络 id | 主机 id

网络 id = 00001000 = 8

主机 id = 00010100.00001111.00000001 = 20.15.1

子网划分允许创建多个逻辑子网，它们存在于单个 A、B 或 C 类网络内。如果没有划分子网，则仅能够从 A、B 或 C 类网络中使用一个网络，这是不现实的。

在一个网络上的每条数据链路必须有一个唯一的网络 ID，在那条链路上的每个节点是同一个网络的一个成员。如果将一个大型网络（A、B 或 C 类）分成较小型的子网，则允许创建互联子网组成的一个网络。那么在这个网络上的每条数据链路都有一个唯一的网络/子网 ID。连接 n 个网络/子网的任何设备或网关都有 n 个不同的 IP 地址，它连接每个网络/子网一个地址。为对一个网络划分子网，使用该地址的主机 ID 部分的一些比特，扩展自然的掩码，产生一个子网 ID。例如，给定一个 C 类网 204.17.5.0，它有一个自然掩码 255.255.255.0，再以这种方式可创建子网：

204.17.5.0：11001100.00010001.00000101.00000000

255.255.255.224：11111111.11111111.11111111.11100000

|sub|

通过扩展掩码为 255.255.255.224，已经从该地址的原主机部分取走 3 比特［由“sub”（划分）指明］，并使用它们构造子网。采用这 3 比特，就可能创建 8 个子网。采用剩下的 6 个主机 ID 比特，每个子网可有多达 32 个主机地址，其中

30个可实际指派到一台设备，原因是为全0或全1的主机id是不允许的（记住这一点是非常重要的）。所以，记住这一点，则创建的这些子网有：

204.17.5.0 255.255.255.224 主机地址范围为1~30

204.17.5.32 255.255.255.224 主机地址范围为33~62

204.17.5.64 255.255.255.224 主机地址范围为65~94

204.17.5.96 255.255.255.224 主机地址范围为97~126

204.17.5.128 255.255.255.224 主机地址范围为129~158

204.17.5.160 255.255.255.224 主机地址范围为161~190

204.17.5.192 255.255.255.224 主机地址范围为193~222

204.17.5.224 255.255.255.224 主机地址范围为225~254

有两种方式表示这些掩码。首先，因为在使用比“自然”C类掩码多3比特的掩码，所以可将这些地址表示为有一个3比特子网掩码。其次，掩码255.255.255.224也可表示为/27，因为在掩码中要设置27比特。

在子网划分的前面所有例子中，注意相同的子网掩码可应用于所有的子网。这意味着每个子网有相同数量的可用主机地址。在多数情形中，为所有子网使用相同子网就结束了浪费地址空间的情况。VLSM是这样一项技术，它允许网络管理员将一个IP地址空间分成不同尺寸的子网，而不是简单的相同尺寸子网划分。本质上而言，VLSM是对一个子网实施子网划分。它也可被描述为，在多个层次上将IP地址分成子网，并依据一个网络上的个体需要分配这些子网。它也被称作无类IP编址。有类编址遵循较早期的IP编址方案，会导致IP地址空间的浪费。

3.2.2 网络和流量：电路和分组（数据报）交换

电信网络是首批出现的网络之一。在电信网络中，用于话音流量传输的技术被称作电路交换。在这种情况中，两个节点之间的电路首先通过一种信令机制建立，信令机制正常情况下使用一个不同的网络，称作一个信令网络，之后这条电路在话音呼叫整个期间保持连接。这种机制的主要优势是可靠性，它在呼叫期间支持连续的流量传输。它也没有为所连接节点之间路由流量而涉及任何额外负担，原因是在呼叫期间，电路路径不会改变。整条消息是作为连续流发送的。就资源使用方面而言，这是低效的。但是，在这种情形中，容量总是得到保障的。

在分组交换中，消息被分段为较小的分组。每条分组标记有源和目的地址及序列号，所以如有必要，可在目的节点处对分组重新排序。之后这些分组在一个复用的和共享的网络上路由，所以网络资源使用是比较高效的。在目的节点处，之后分组按序被重组。在分组被发送的同时，不同分组可走到目的节点的不同路径。

在分组交换网络中，使用电路交换的一种模拟法，称作虚拟电路交换。在这种情形中，在分组传输之前，要提前建立连接，之后分组按序被传递。

3.2.3 网络设备

如前所述，一个网络由通过许多流量转发设备连接的多台主机组成。这些设备可工作在一层或多层，从层1到层3。

最早这样的设备之一是一台以太网络集线器或一台重发器式集线器，工作在层1或物理层。这些被用来使用多条双绞线以太网线缆将多台主机连接在一起。这些使被连接的设备行为就像连接到单个网络分段。集线器不管理任何流量，但来自一个端口的任何分组被复制，之后在所有其他端口上重新广播。这种不复杂的设计意味着，随着连接到集线器的主机数增加，网络分段利用率变得低效。许多集线器也可通过独立端口实现堆叠，但网络冲突方面的增加使这种方法不是一项非常具备可扩展能力的解决方案。

这些集线器可检测过量的冲突，并可隔离特定端口。相比在一条多段以太网线缆上简单地连接设备的方法，基于集线器的网络是比较具有扩展能力的。通过一个集线器连接的每个网络分段被称作一个冲突域。集线器没有读取和截获通过它们的任何分组的任何内容的能力。如今这些设备几乎被淘汰，且事实上被终止使用了。集线器的多数功能可由交换机复现，它们是更智能的层2设备。

交换机的前辈是网桥。这些设备工作在层2或数据链路层。网桥具有读取连接到它们的各设备MAC地址的能力，并可实施基本的学习，简单地将帧仅发送到拟设的目的设备。本质上而言，网桥隔离冲突域，并具有拓扑学习能力，仅将到达流量复制到它拟设的那些端口上。当连接多台网桥时，极端重要的是，要避免带有回路或循环的端口之间的路径，这可能导致整个网络的广播风暴和性能降级。出于这个目的，网桥采用生成树协议（STP）。经典的STP已经大部分为快速生成树协议（RSTP）所替换。

一个广播风暴是这样一种状态，其中当在一个环路中连接的以太网交换机中没有启用生成树时，未知单播和广播分组无休止地循环，导致硬件变得过载。一棵生成树被用来防止以太网上的环路。生成树阻塞一个接口，所以去除了环路和广播风暴的可能性（见图3.6）。

一次广播风暴可逐渐地变得越来越严重，原因是在没有到达它们的目的地之前，分组继续无休止地循环流动。它们很快由新的广播加入，这些新的广播也陷入循环之中。最后交换机将变得流量不堪重负，崩溃或重启。交换机在一次风暴期间变得如此过载，以致任何形式的管理都变得不可能。在风暴开始之后，小型低端交换机在性能严重地被影响之前可能需要一两分钟。在不同型号产品间所花费的时间似乎是不同的。由于它们较高的速度和暴力转发能力，大型高端交换机几乎立刻就受到影响。

交换机是或多或少替换了网桥的设备。交换机具有将网络分成多个冲突域的能力以及学习网络拓扑的能力。但这些也为所连接的设备之间提供了全双工路径，以

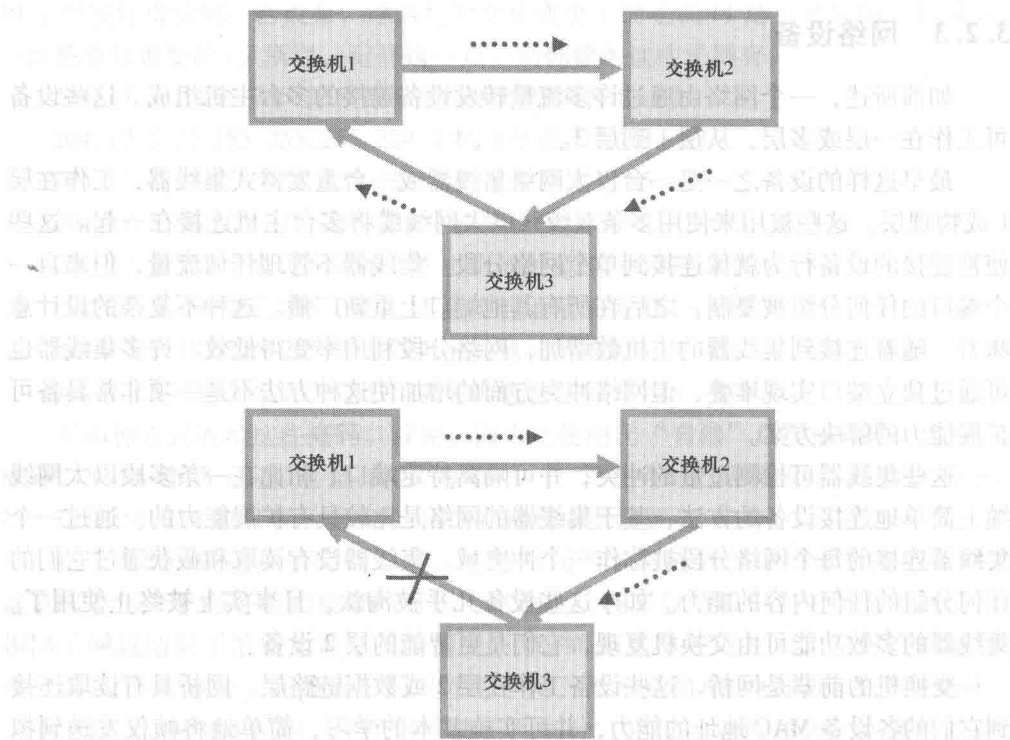


图3.6 广播风暴和STP

便防御设备间的冲突。正常情况下，这是通过使用一种内部转发机构完成的，该机构在两个方向转发网络流量，比一个个体网络接口要快得多。

一旦交换机学习到网络拓扑，则通过使用存储和转发、直通、不分段（Fragment-Free）或自适应交换等，实施流量转发。

1）存储和转发交换：在将帧转发到其他端口之前，交换机缓冲并验证这些帧。这是固有的最慢方法。

2）直通交换：在这种情形中，即刻实施错误检查。交换机仅读取帧的硬件地址，并开始转发它。在网络拥塞的情形中，交换机退回到存储和转发过程。

3）不分段交换：在这种过程中，如果在帧的前64字节中检测到冲突，那么因为错误而丢弃帧，不会转发。需要由端设备进行帧错误检查。

4）自适应交换：在这种情形中，交换机可自动地在多个节点之间交换，取决于流量情况。

路由器是一种层3设备，它连接两个或多个不同网络。这些网络也可以是非IP网络，因为层3路由器的主要功能之一是协议转换。路由器具有读取源和目的地址的能力，这些地址被用来构建路由和转发表，使用这些表从正确的外发接口朝目的网络进行流量转发。如前面看到的，一个网络由多台路由器组成，它们使用路

由协议交换有关目的地址的信息。在这种信息交换的基础上，每台路由器构建其路由表，指明在互联网中任何两个系统之间的首选路由。一个大型计算机网络可能被分成多个较小的和更可管理的网络（这些网络经常反映出一个组织的细分情况），并由路由连接起来。这种被细分的网络称作子网，本质上是设备或主机的逻辑分组。一台路由器经常工作在两个逻辑平面，即控制平面和转发平面。这类似于电信网络中信令和电路交换网络中的分隔（Division），例外是这些平面是使用一组共同的接口集与对端路由器中相应平面交换分组的。控制平面负责更新路由表（对应于网络的动态变化的拓扑），而转发平面在进入和外发接口之间转发分组。小型办公室家庭办公室（SOHO）商务或应用使用单台交换机或路由器，可能连接到一台宽带接入设备，如一台DSL或一台调制解调器。但是，路由器，像恐龙或袋鼠一样，有所有尺寸的，小型的、大型的到非常大型的，取决于性能、流量容量及其在网络拓扑中的位置。像思科CRS-1的高端路由器由各ISP使用与其他ISP互联，原因是这些可用于管理大型流量流（Flow）。

最常见的是，用在各种企业、ISP或组织中的路由器可被分类为接入、分配或核心路由器。接入或边缘路由器是用在接入ISP网络的那些路由器，范围涉及家庭路由器或SOHO路由器。这些路由器中一些路由器甚至可基于开源的基于Linux的通用现成（off-the-shelf）商用硬件的。分配路由器扮演来自多台接入路由器之流量汇聚器的角色。它们也作为流量整形器和服务质量（QoS）实施器（Enforcer），确保不同类型的流量，如因特网协议上的话音（IP电话，VoIP）、多媒体或数据传递，得到合适的资源分配。其主要功能是在不同的主网络区之间提供高带宽连接能力，同时交换来自各分配或接入路由器的汇聚路由信息，以便构建路由表（在不同区路由器间进行交换）。这些路由器主要运行BGP。

此外，路由器可被分类为边缘、用户边缘、提供商间边界或核心路由器。这些是更面向ISP的路由器分类。边缘路由器处在ISP网络的边缘，并与其他ISP或大型企业路由器交换信息。用户边缘路由器主要是针对个体或SOHO用户流量的汇聚器。在这种情形中，核心路由器，作为网络的骨干，与各边缘和分配路由器交换流量和路由。层1 ISP提供的绝大多数路由器及其互联网络，作为网络的骨干。该比喻就像在星系的中心处有一个或多个“黑洞”的星星的银河系分布，银河系的各分系有不同的星系分布。最近时间以来，因特网上流量的特征发生了剧烈变化，多媒体流量构成总流量的相当大的比例，这使路由器带有有效的QoS实现成为必要的。

3.2.4 网络流量路由

基本上而言，路由是在沿所选中路径转发流量的过程，目的是最小化一个网络中的成本。因为正讨论分组交换网络，所以路由指引导分组转发，从源节点到目的节点，沿最小开销的路径。在分组转发的操作中涉及各种设备，包括路由器、网

桥、交换机和防火墙。

在路由中涉及的主要设备是路由器或层3交换机。路由器维护路由表，这些表用于引导流量沿到目的地节点的路径到下一节点。分组的目的地址作为查找路由表的一个索引。路由协议活动的主要部分涉及作为对变化的状态的响应，维护最新的路由表。

在分组交换网络中，路由意味着转发分组，从其源到目的地，通过任何中间节点。这是通过前面提到的一台或多台设备完成的，如网桥、交换机、路由器等。虽然任何现成商用硬件可被用来开发一台交换机或一台路由器，但更常见的由于性能原因，多数情况下，这些设备使用专用的或专门的硬件，如应用特定的集成电路（ASIC），执行分组转发。

路由背后的主要思路是逐跳地转发分组。正常情况下，在一个局域网（LAN）中，分组被转发到一台网关路由器，它维护一个转发表，之后使用转发表确定分组要被转发到的下一个目的地。

为理解路由，需要考虑的两个主要概念是路由表和路由协议。一个路由表或路由信息库（RIB）是一个数据结构，虽然概念上像一个表，但实践中是使用前缀树实现的。本质上而言，路由表反映路由器对网络拓扑的当前理解。路由表是由路由器软件构造的，以路由协议使用信息广播或组播完成的。在静态路由的情形中，路由表项是人工更新的。在现代路由器中，由路由表中的信息构造一个较小的转发表，之后该表用于实际地转发分组。

1. 路由表

一个路由表主要由如下字段组成：

1）目的网络标识符。

2）对于分组而言，路径的成本（依据某种度量构建的）。

3）要将分组转发到的下一个节点。

在一些情形中，各表项也许包括额外信息字段，如访问列表信息和与路径相关联的QoS。

如可从图3.7中看到的，网络路由表输出包含有关如下方面的信息：

1）网络目的地和网络掩码：确定网络标识符。

2）网关：包含有关要转发分组之下一跳的信息。

3）接口：确定外发接口，在该接口上可访问网关。

4）度量指标：是与路径相关联的成本度量（对于路由协议特定的成本度量指标）。

2. 路由协议

本质上而言，一个算法是被设计用来控制一个过程的规则集。在控制各种内部因素和外部因素的同时，一些路由算法也支持路由分组的多条可选路径，称作多路径路由。网桥也转发帧，但在网络的有限邻域内转发而已。路由使用IP地址，它

```
Command Prompt
Copyright (c) 2009 Microsoft Corporation.  All rights reserved.

C:\Users\dmiruke>
C:\Users\dmiruke>netstat -rn
===========================================================================
Interface List
 18...06 25 d3 c8 39 b4 ......Microsoft Virtual WiFi Miniport Adapter
 12...90 e6 ba a0 aa 48 ......Atheros AR8132 PCI-E Fast Ethernet Controller (NDIS 6.20)
 11...00 25 d3 c8 39 b4 ......Atheros AR9285 Wireless Network Adapter
  1...........................Software Loopback Interface 1
 28...00 00 00 00 00 00 00 e0 Microsoft ISATAP Adapter
 13...00 00 00 00 00 00 00 e0 Microsoft 6to4 Adapter
 41...00 00 00 00 00 00 00 e0 Microsoft ISATAP Adapter #2
 22...00 00 00 00 00 00 00 e0 Microsoft ISATAP Adapter #4
 26...00 00 00 00 00 00 00 e0 Teredo Tunneling Pseudo-Interface
 27...00 00 00 00 00 00 00 e0 Microsoft ISATAP Adapter #3
===========================================================================

IPv4 Route Table
===========================================================================
Active Routes:
Network Destination        Netmask          Gateway       Interface  Metric
          0.0.0.0          0.0.0.0    192.168.100.1  192.168.100.41     25
        127.0.0.0        255.0.0.0         On-link        127.0.0.1    306
        127.0.0.1  255.255.255.255         On-link        127.0.0.1    306
  127.255.255.255  255.255.255.255         On-link        127.0.0.1    306
    192.168.100.0    255.255.255.0         On-link   192.168.100.41    281
   192.168.100.41  255.255.255.255         On-link   192.168.100.41    281
  192.168.100.255  255.255.255.255         On-link   192.168.100.41    281
        224.0.0.0        240.0.0.0         On-link        127.0.0.1    306
        224.0.0.0        240.0.0.0         On-link   192.168.100.41    281
  255.255.255.255  255.255.255.255         On-link        127.0.0.1    306
  255.255.255.255  255.255.255.255         On-link   192.168.100.41    281
===========================================================================
Persistent Routes:
  None

IPv6 Route Table
===========================================================================
Active Routes:
 If Metric Network Destination      Gateway
 26     58 ::/0                     On-link
  1    306 ::1/128                  On-link
 26     58 2001::/32                On-link
 26    306 2001:0:4137:9e76:849:b78:3f57:9bd6/128
                                    On-link
 11    281 fe80::/64                On-link
 26    306 fe80::/64                On-link
 26    306 fe80::849:b78:3f57:9bd6/128
                                    On-link
 11    281 fe80::9ce0:17f8:27d1:94e3/128
                                    On-link
  1    306 ff00::/8                 On-link
 26    306 ff00::/8                 On-link
 11    281 ff00::/8                 On-link
===========================================================================
Persistent Routes:
  None

C:\Users\dmiruke>_
```

图 3.7　一种典型的路由表输出

是有结构的，而桥接使用无结构的地址，即 MAC 地址。桥接（透明的或学习型网桥）指多个网络分段在层 2 由称作网桥的设备进行连接。交换机和网桥指类似设备，例外是，交换机是多端口网桥。因为桥接简单地洪泛各分组，而交换机使用 MAC 地址学习方法，将 MAC 地址与到达和外发端口相关联，对于被转发帧中源和目的地址的格式不做任何假定。

有结构地址支持单个路由表，为从一个网络到另一个网络的分组高效路由而加以维护。在桥接和交换用在局部化环境内时，路由用于在大距离上以及从一个域到另一个域转发分组。

当讨论在一个网络中路由流量时，需要考虑如下问题：

1）基础网络提供哪种类型的服务（ToS），即网络携带的流量特征是什么？

2）网络的协议栈是什么样子的？

3）路由器元素的设计，即路由器用来管理和路由流量的配置和处理能力有哪些？路由器的目的是为转发用户流量而计算最佳路径，以及检查用户流量流

(Stream)(如有必要),并实施额外所需的处理。

4)网络使用什么类型的拓扑,以及网络拓扑变化得有多快?

流量管理方面是如何在网络中处理的?由路由器转发的用户流量被称作数据平面流量,而发生在路由器间和路由器内组件内部的信息交换被称作控制平面流量。网络配置等是由所谓的控制平面流量处理的。

在一个典型网络拓扑(见图3.5)中,单个路由协议在一个典型网络或子网内的所有路由器上执行。本质上而言,一个路由协议是一个分布式算法,它将网络拓扑更新分发到各路由器,并为每个节点计算它的路由表版本。

3. 路由表分类

根据特定拓扑,该算法的适用性和效率(efficacy),存在各种路由算法,用在各种网络拓扑中。根据各种属性,对路由算法进行分类:

1)根据算法本质上是静态的还是动态的,实施分类。在静态算法的情形中,本质上路由表是在所有路由器上以手工方式传播的。在较小型网络中或在层1 ISP间的交换点中,情况是这样的,其中网络路由不会非常频繁地变化。在动态路由算法的情形中,取决于所采用的路由算法,通过网络拓扑改变广播或组播,更新路由表。

2)根据网络拓扑是扁平的还是层次结构的,也可进行分类。

3)在路由协议中,经常见到的一个区别是在内部协议和外部协议之间做出的。内部网关协议(IGP)被用于在属于单个AS或一个路由域的各路由间进行流量路由。一个AS本质上定义属于单个管理域(或以比较简单的术语来说,一个组织拥有的路由器集合)的一个网络。正常情况下,一个组织可拥有在不同位置的多个网络,但这些将仍然被看作单个AS。为从一个AS内到达另一个AS,采用外部网关协议(EGP)。

4)路由协议也使用各种计算属性进行分类,如跳计数、所用的距离度量指标等。但是,这是不常用的方案。

5)路由算法的主要分类之一是基于在路由协议中所用算法的类型。一些主要算法有距离矢量(DV)、链路状态、混合DV和链路状态以及路径矢量。随着讨论的进行,将比较详细地解释这些算法。为做到直接参考,将这些分类收集到表3.1中。其中,IGRP表示内部网关路由协议;RIP表示路由信息协议;EIGRP表示增强内部网关路由协议;OSPF表示开放最短路径优先。

表3.1 路由协议分类

分类属性	分类	协议
控制	静态	手工配置
	动态	多数协议
拓扑	扁平的	IGRP、RIP、EIGRP
	层次结构的	OSPF、BGP

（续）

分类属性	分　类	协　议
范围	内部（域内）	RIP、IGRP、OSPF
	外部（域间）	BGP
路由计算参数	跳计数 带宽 时延 可靠性 负载	RIP、IGRP、EIGRP
基本路由计算算法	DV	RIP、IGRP
	链路状态	OSPF
	混合（DV + 链路状态）	EIGRP
	路径矢量	BGP

就尺寸、规模和携带的流量总量而言，公共交换电话网（PSTN）仍然是世界上最大的网络。虽然 PSTN 正在各种场合得到 IP 网络的增强、替换或补充，同时虽然这是一部有关全 IP 网络的书籍，但有关 PSTN 的一个简短讨论是适宜的。PSTN 中的主要用户流量是用户话音呼叫，一定程度上是数据呼叫。PSTN 使用一种不同的层次结构编址（即编号或 E.164）方案。IPv6 编址方案的一些元素要将其存在归功于 PSTN 编址方案。多数上述考虑也适用于 PSTN。在 PSTN 中，控制和管理平面流量是在一个独立的信令网络中传输的，而用户流量（话音和数据呼叫）通过一个主要是消息交换的网络传递的。

针对单播、广播、组播和任意播消息，存在多个路由协议基础算法。单播消息被交付到单个目的地。广播消息被发送到一个网络中的所有节点，而组播消息被发送到为该组专门注册的一组节点。任意播消息被转发到节点组的任意一个节点，正常情况下是最接近源的一个节点。将主要焦点放在单播和组播路由算法和协议上。

一个路由表可包含六种路由之一，如表 3.2 所示。

表 3.2　一个路由表中的路由类型

路由类型	描　述
主机路由	到一台特定主机而不是一个网络的路由。一个掩码是类型/32或 255.255.255.255
子网路由	一个主网络的一部分。一个子网掩码用来确定子网，如 10.10.1.0/24（带有掩码 255.255.255.0）

（续）

路由类型	描　述
汇总（子网的组）	单条路由指代一组子网。10.10.0.0/16（255.255.255.0）为带有前缀10.10.0.0的所有网络提供了一个汇总
主网络	一个有类的网络，带有一个原生掩码（带有掩码255.255.0.0的10.0.0.0/8）
超网（主网络的组）	单条路由指代一组主网络。10.0.0.0/6指代10.0.0.0/8和10.0.0.0/16网络
默认路由	这被显示为0.0.0.0，当目的IP地址没有匹配路由表中的任意前缀时，用于转发分组

图3.8所示是一个假想网络的例子，表3.3所示是相应路由表的一个例子。

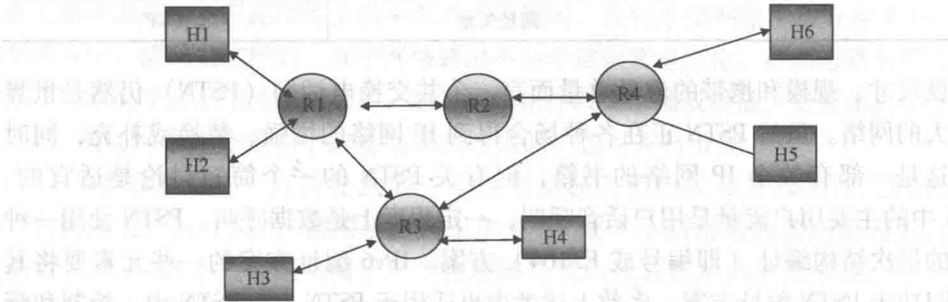


图3.8　解释路由表结构的一个网络

表3.3　图3.8中网络的路由表

在R1处的路由表		在R2处的路由表		在R3处的路由表		在R4处的路由表	
目的地	外发接口	目的地	外发接口	目的地	外发接口	目的地	外发接口
H1	直连	H1	R2-R1	H1	R3-R1	H1	R4-R2
H2	直连	H2	R2-R1	H2	R3-R1	H2	R4-R2
H3	R1-R3	H3	R2-R3	H3	直连	H3	R4-R3
H4	R1-R3	H4	R2-R3	H4	直连	H4	R4-R3
H5	R1-R2	H5	R2-R4	H5	R3-R4	H5	直连
H6	R1-R2	H6	R2-R4	H6	R3-R4	H6	直连

对于小规模网络，采用恒定的或相对静态的拓扑，路由表可由管理员提前指定，网络可使用静态的、非适应性路由。对于大型网络和动态拓扑，则使用动态的、自适应的拓扑是必要的。在自适应路由中，通过检测网络拓扑中的变化，由路由协议自动地构造路由表。一些主要的自适应路由协议是RIP、OSPF和一些准专用协议，如中间系统到中间系统（IS-IS）、IGRP和EIGRP。OSPF和一些派生的协

议是因特网中的主要协议。

在检测到网络拓扑中的变化之后，多数自适应路由协议使用各种算法计算路由路径。比较占主导地位的算法中的三个算法和从这些算法中派生得到的协议类有 DV、链路状态和路径矢量协议。

路由的最早方法之一涉及 DV 算法。在这种算法中，一个成本或一个数值被指派到节点之间的每条链路，分组沿从源到目的地的最小成本的路径发送。正常情况下计算的总成本作为与每条链路相关联的成本之和。在实现中，与每条链路相关联的成本是以从下一个节点到达目的节点所需的跳数指定的。

当节点启动时，它仅知道其邻居和与到达每个邻居相关联的成本。路由表中的每个表项由下一跳节点的地址和到达该节点的总成本组成。每个节点向每个其他邻居发送周期性更新，这些更新是有关到达目的节点的信息的。邻接节点分析路径和接收到的关联成本，并将这些成本与以前已知的成本进行比较。如果新成本较小，那么新路由就被安装在路由表中。当任何节点宕机时，使用这个节点作为下一跳的所有节点，从其路由表中清除关联的表项。

RIP 和 IGRP 使用 DV 算法有一些变化。也存在使用这个算法的各种专门的自组织网络（Ad Hoc Network）路由协议。非常常见的是，将看到链路状态和 DV 算法的一个组合体实现为一个路由协议的组成部分。正常情况下，这被看作专门网络路由方案的组成部分。

另一种主要算法是链路状态算法。在这种情形中，网络中每个节点维护它所驻留网络的一个“映射”。该“映射”本质上由一个图表示。每个节点通过洪泛向所有其他节点发出链路状态信息。此后，每个节点独立地计算从自己到所有其他节点的到目的地的路径。对此，各协议使用 Edgar Dijkstra 的最短路径优先（SPF）算法。这个算法由 OSPF 和在有线与自组织无线域中派生的协议所使用。

在因特网和路由中使用的主要概念之一是 AS 的概念。对于不同路由协议，一个 AS 指代不同事物。对于 OSPF 和 EIGRP，一个 AS 仅是在单个行政管理下的一组地址，首选带有单个前缀或一组受限的前缀。但是，各路由将不做交换，即使在单个行政管理下的各域内部也是如此，除非显式地加以指明的情况。也存在一些保留的 AS 号，是为中转 ISP 网络指派的。所有 AS 号都是唯一的。64512 ~ 65535 的号段是为私有用途保留的，即这些号段不应在因特网上通告。正常情况下，因特网编号管理机构（IANA）是指派 AS 号码的一个机构。

使用 DV 和链路状态算法的各协议主要是用在各 AS 内部的 AS 内或域内协议。但是，如果这些协议用于在不同 AS 之间路由流量，则路径计算变得不可行，原因是所涉及网络的大型尺寸造成的。DV 算法变得不可行，是由于巨大数量的跳数导致的，而链路状态算法要求大量内存和计算能力以及带宽［用于链路状态通告（LSA）洪泛和更新］是由于巨大数量的节点导致的。

所以对于域内路由的目标，使用路径矢量协议。在路径矢量中，假定每个 AS

由一个或多个节点的集合表示，这些节点通告那个域或 AS 的路由表到其他域。这些协议类似于 DV 协议，例外是仅有代表节点将发出该 AS 的更新。在信息交换过程中，各节点在 AS 中交换路径。虽然路径矢量算法以类似于 DV 算法的方式工作，它不通告一个度量指标，而是通告目的节点和通过该 AS 到目的节点的路径。与路径一起，也交换各种路径属性。各 IGP 工作在一个 AS 内。IGP 协议包括 RIP、IGRP、EIGRP、OSPF、IS-IS（Deepankar Medhi）等。主要要求是在一次拓扑改变之后，这些算法要快速地重新计算高效路由。各 EGP 是工作在各 AS 之间的那些路由协议。主要的 EGP 有 EGP 和 BGP。对一个 EGP 的主要要求是当从一个 AS 将路由通告到另一个 AS 时，它要指定各种复杂的路由策略并汇聚路由信息。

正常情况下，各 IGP 和各 EGP 协作，确保从因特网中的任意一个点到任意其他点都存在端到端连通能力。但是，这是以路由策略和通告方面的严格规则做到的。也存在一个已定义的管理距离（AD）度量指标，它指定被通告路由路径被信任的程度。正常情况下，通过遵循严格的路由策略，一个 AS 将汇聚的一个路由集注入另一个 AS。

见表 3.4 和表 3.5。

表 3.4　路由表比较

	DV	高等 DV	链路状态	路径矢量
	RIP、IGRP	EIGRP	OSPF	BGP
扩展性	低	高	良好	优
带宽	高	低	低	低
延迟	低	中等	高	高
CPU（中央处理单元）使用率	低	低	高	中等
收敛	慢	快	快	中等
配置	容易	容易	中等	配置
VLSM	否	是	是	是
多路径支持	否	是	是	否
无 IP 支持	否、是	是	否	否

表 3.5　路由协议汇总

协议	算法	IETF 标准	内部/外部	更新	度量指标	VLSM/CIDR 支持	传输协议	汇总
RIPv1	DV	是	内部	30s	跳数	否	UDP	自动的
RIPv3	DV	是	内部	30s	跳数	是	UDP	自动的
IGRP	DV	否（思科）	内部	90s	复合的	否	UDP	自动的
EIGRP	高等 DV（双重的）	否（思科）	内部	触发的	触发的	是	RTP	自动的/手工的

（续）

协议	算法	IETF 标准	内部/外部	更新	度量指标	VLSM/CIDR 支持	传输协议	汇总
OSPF	LS	是	内部	触发的	成本	是	IP	手工的
IS-IS	LS	是	内部	触发的	成本	是	IP	自动的
BGP	路径矢量	是	外部	增量的		是	TCP	自动的

注：RTP—可靠传输协议；UDP—用户数据报协议；IETF—因特网工程任务组。

与路由有关的一些其他特定专题包括：

1）IP TOS字段：IP中的TOS字段被用来分类流量流（Stream）并对其排出优先级。定义了5个值：正常、最小成本、最小时延、最大吞吐量和最大可靠性。但是，在当前，多数路由协议，包括IPv6，不支持基于ToS的流量优先级。如果要使用，则将有必要维护多个路由表，这取决于ToS值。

2）IP选项——严格源路由：这指定了分组应该走的准确路径，并包括沿路径的每一跳。由于IP首部的尺寸限制，最大跳数要小于9。松散源路由指定沿路的一些跳。正常情况下，这些选项用来对网络中的路径进行排错。在IP首部中有额外选项，也对路由器造成附加处理开销，所以这些是相对稀少地使用的。

3）多路径路由：对于流量的负载均衡，这指定了多条可选路径。像EIGRP的一些路由协议使用这个选项。

4）默认路由：当没有其他路由可由分组使用进行转发时，这是要使用的最不具体的可能路由。可手工地配置或动态地学习。

5）默认网关：当不支持IP路由（即当一台交换机处在网桥模式时），一台默认网关被指定为一个特定的IP地址，能够将分组发送到除自己以外的网段。

6）默认网络：这是由诸如IGRP和EIGRP的协议所使用的，将分组路由到默认网络分段。

4. 选择或设计一个路由协议的核心考虑

1）路由层次结构：当网络中的路由器数量增加时，这也增加了控制流量所需的带宽，以及为路径计算所需节点上的处理器负载，特别在链路状态算法中更是如此。这是许多协议（像OSPF）支持一种层次结构网络配置的原因。

2）路由计算：在链路状态协议中，使用网络拓扑知识，每个节点计算从自己到网络中每个其他节点的最短路径。这可能是非常处理器密集（即需要大量计算）的，特别当网络中的路由器数非常庞大时更是如此。这是网络以一种层次结构形式将不同区结构化配置OSPF路由协议的原因之一。

3）路由器信息流：多数路由协议使用周期的或触发式的更新，将网络拓扑的变化传递到邻接节点。一些协议交换完整的路由表，而一些协议仅交换变化的信息。当网络中的路由器数非常庞大时，控制流量可能成为总体流量的一个相当大的

部分，特别当更新频率较高时更是如此。这延迟了路由表更新，并可能导致在网络不同部分中各节点之路由表的某些不一致。当一个网络拓扑快速变化时，这可能导致网络为不稳定的。

4）路由路径选择：路径选择涉及与一种特定算法组合使用，应用某种路由度量指标，为转发分组而选择最佳路由。度量指标可能涉及各种参数，如带宽、网络延迟、路径成本、负载和链路可靠性等。

5）收敛性降低：在一些情形中，网络拓扑变化是非常频繁的，特别在自组织无线网络中更是如此。重要的是，对所有节点，在网络拓扑再次改变之前，网络路由表反映最新的改变。多数情况下情况未必如此。在收敛背后的思想是，所有节点都有网络拓扑的一个共同视图。当不同节点有网络拓扑的不同视图时，对网络流量可能发生非常严重的事情，包括流量丢失、由于带有环路的路径上IP数据报的重复重发导致的网络拥塞等。最小化收敛意味着，使所有节点形成网络拓扑改变（在网络拓扑中的任何改变之后）的一个共同视图所需的时间应该最小。

6）路由汇聚：路由汇聚是在给定存在一条特定路由的情况下，产生一条更通用路由的方法。路由汇聚也由大型区域网络或AS使用，降低到处传递（Pass Around）的路由信息量。采用仔细地将网络地址分配给客户端的方法，大型网络可仅将一条路由通告到区域网络而不是数百条路由。

7）扩展性：对一个较小型网络工作的方法，对具有更多节点的一个网络，可能并不工作。对于一个较小型的网络，RIP工作良好，但对于具有10倍现有节点的一个网络而言，路由器控制流量开始占据网络流量的一大部分，在任何拓扑改变之后可能影响收敛。

8）鲁棒性：这意味着，在计算机系统中，在不管环境中或输入中异常事件情况下，硬件和软件继续工作的能力。通过采用变化的负载特点和模糊测试（即改变输入数据向量，产生各种类型的错误数据），使用故障评估，测试产品的鲁棒性或可靠性。

5. 路由算法的比较

DV协议配置和部署是简单的，但对于大型网络扩展性不好，这是由于网络的有限直径造成的，这些协议要解决缓慢收敛性质。因为这些协议基于一个跳计数度量指标，而不是链路状态，这些协议忽略其他因素，如链路的带宽利用率、速度等（其他链路状态算法中考虑这些因素），所以可能给不出真实的路由度量指标。

这些缺陷在诸如OSPF和IS-IS的链路状态协议中以及诸如EIGRP的一种无环路的DV协议（思科专有的协议）中得到了解决。

要考虑的另一方面是路径选择，即使用路由度量指标从许多可能路由中计算最佳路径。正常情况下，人们期望路由协议使用诸如带宽、网络延迟、跳计数、路径成本、负载、链路可靠性、最大传输单元（MTU）等计算一个成本度量指标。各种路由协议使用不同的启发式方法从邻居节点中学习到的许多路径中选择最佳路

径，并将之存储在链路状态或拓扑数据库中。

同样，一个网络可从能够携带中转流量的另一个网络中接收路由。在这种情形中，各网络可运行不同协议，需要一种处理方式，对由一种路由协议通告的路径将其置信度水平高于另一种路由协议。出于这个目的，形成了 AD 的一个概念。这类似于置信度度量或水平，其中一个较小的 AD 指明这些路由是从一个比较可靠的路由协议学习到的。一名管理员可指定具体的静态路由，以便能够调试一个路由问题或路由表。

如在前面看到的，因特网由多个层次结构的 AS 组成。一个 AS 可对应于一个 ISP，或一个 ISP 可管理多个 AS。AS 层次的路径由 BGP 选择，它是一种外部路由协议。BGP 将接收各种 AS 或域之间的多条路径。但是，从总的可能路径中选择单条路径或路径的子集，不仅取决于成本度量指标，而且取决于不同组织的商务政策，条件是控制不同于交换路径的情况。出于这个原因，像 BGP 的路由协议也需要支持政策机制，采用这种机制，管理员可建立层次结构政策，以便为各域或 AS 之间路由流量而选择“最佳”路径。多数情况是，商务政策意味着互惠的程度，其中允许其他域传递中转流量。从成本角度看，这并不总是得到最优路径。

网络拓扑的类型在路由和流量处理中扮演一个重要角色。需要考虑的一些因素有网元连接的大小和规模、网络内各节点的处理能力等。例如，具有有限节点基础设施网络的一个有线网络所需的路由表，相比动态的无线网络，具有非常不同的特征，在无线网络中，网络拓扑变化是非常频繁的，且各设备可能具有低得多的处理能力，并可能具有诸如电池电量限制等约束。网络的整体连接能力也扮演一个角色，即一个整体的网状连接网络和一个层次结构网络中的路由算法（是不一样的）。

6. 路由度量指标

路由度量指标被用来确定最优路径。但是，许多度量指标的值是由各种因素控制的，这些因素可分类为环境的和网络相关的因素。环境的因素是不受网络反馈影响的因素，如节点的放置和移动性、节点的属性等。网络相关的因素被定义为直接或间接地取决于网络中流量的那些因素，如拥塞、由流量间和流量内流导致的干扰以及网络拓扑。

度量指标也可表征为

1）组合的度量指标，是从其他度量指标以数学方式组合得到的，如链路成本，是从时延、带宽、流量水平等计算得到的。

2）OSI 栈的层，提供计算度量指标所需的信息。虽然传统上来说仅使用网络层测量数据，但如今，采取一种不同的方法，其中在定义路由度量指标时，要考虑跨层交互。

3）一些路由度量指标是以分析方式推导得到的，如带宽，同时一些路由度量指标是从经验测量数据得到的。

4）与路由度量指标有关的信息可以各种方式得到：

① 节点相关的信息：在没有太多付出的情况下，直接从节点得到信息，如节点的接口数、通信成本等。

② 被动监测：通过观测来自一个节点的进入流量和外发流量收集信息。

③ 捎带探测：通过在常规流量流或控制流量中探测信息实施测量。

④ 主动探测：为度量一条链路的性质产生特殊分组。

7. 路由分析

一个相对新的领域正在出现，称作路由分析，它提供 IP 云内路由行为的可视化能力。一组指定的路由器（DR）作为数据收集代理，与各 L3 级路由器建立邻接关系，并被动地侦听各路由器交换的控制平面消息。即使这些系统主动地参与到控制平面，它们也不影响路由或流量流。各路由器也必须支持正部署在网络中和正被测量的所有路由协议。

路由分析系统是这样工作的，与一个层 3 网络中的单台路由器建立一个关系（邻接关系），接着是被动地侦听正在由路由器交换的控制平面消息。通过成为控制平面的组成部分，路由分析系统实际上作为一台被动路由器，具有其他网络路由器的相同路由知识，但没有转发实际数据分组的能力。虽然系统正主动地参与控制平面，但它们不能影响数据是如何在网络各处路由的。各系统也必须支持各种路由协议，以便实际上分析由网络路由器通告的更新。

由此，路由分析系统可即刻了解控制数据流和网络中发生的所有事件，如流量链路的倒换（Flapping）。使用这种数据收集法和各种数据分析技术，则可能观察到网络行为模式，甚至检测到网络故障的一种新类型，即使这些故障是间歇性的也可做到这一点。

基本上而言，路由分析系统提供：

1）一个网络拓扑地图的自动发现和构造。

2）IP 网络云的实时可视化能力。

3）以实时方式对网络拓扑之故障和不稳定性的监测、分析和可视化。

4）故障检测（或甚至预测）和纠正所需时间的降低；路由数据和事件与路由（和应用）性能的相关。

5）即使是影响性能的异常路由事件、故障或协议异常的检测。

一些公司，如 Packet Design、Netcordia、Solana 网络和 Iptivia，都提供路由分析解决方案。

8. 路由器组件和架构

一台路由器由如图 3.9 所示的基本组件组成：

1）多个网络接口卡（或线卡）与附接网络接口。

2）多个处理模块或监控（Supervisor）卡。

3）内部交换结构。

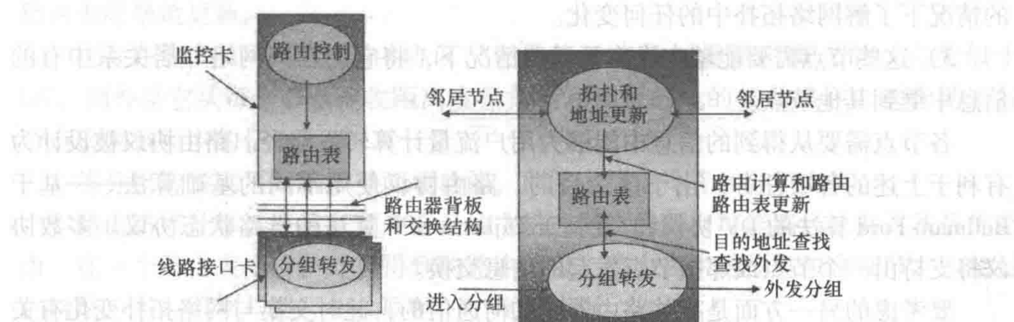


图3.9 一台通用路由器的基本组件和架构

为做到比较快速地处理，最常见的是，这些组件是使用一个或多个ASIC构造的。在到达接口接收分组之后由处理模块处理，然后通过交换结构转发到外发接口。其他功能，如接口配置、统计数据收集等，是在管理或控制平面中处理的。

逻辑上而言，路由器架构或功能可分成两个组件（经常称作平面）。这些平面如下：

1）控制平面：这个平面负责与其他路由器参与交互，执行路由协议，接收和广播（或单播或组播，取决于该协议）更新的网络拓扑信息，并构建或更新路由表。同样各种其他功能，如丢弃分组或分组的QoS服务，也是在控制平面中实施的。

2）数据和转发平面：这个平面实际上负责转发流量。来自路由表的信息被用来构建转发表，该表实际上用来将分组转发到外发的目的地。正常情况下，图3.9中的监控卡实现控制平面，而接口或线卡实现数据平面。多数时间，线卡使用三路内容可寻址内存（TCAM）进行转发路径的快速存取。

3.3 路由协议

如早期在编程方面学到的，程序=算法+数据结构。这里使用同一模型，在深入地理解路由网络流量过程之前，需要得到这三方面的一般理解。在讨论路由和相关问题的整体观点之后，需要将焦点放在分组路由中实际上涉及哪些方面。

需要从三个不同观点讨论路由：协议、算法和数据结构（保存路由更新），如路由表和转发表。

注意，路由协议实现的是运行在多个节点（或者在这种情形中是路由器）上的本质上的分布式算法，即处在分布式环境中。

这里以经验方式讨论这个问题：

1）有多个或分布式的节点，假定它们在网络间转发或中继用户流量。

2）明显地，各节点将需要有关网络拓扑状态的信息，也需要在没有任何延迟

的情况下了解网络拓扑中的任何变化。

3）这些节点需要能够在没有延迟的情况下，将它们在其网络邻居关系中有的信息中继到其他节点。

各节点需要从得到的信息中能够为用户流量计算转发路径。路由协议被设计为有利于上述的各项任务。出于这个目的，路由协议使用不同的基础算法——基于 Bellman-Ford 算法的 DV 协议和/或基于 Dijkstra SPF 算法的链路状态协议。多数协议将支持由一个节点或邻接节点发起的信息交换。

要考虑的另一方面是有关路由器是如何通信的，此时交换与网络拓扑变化有关的信息。

自因特网诞生之前，在电信世界中就使用的两个术语是带内信令和带外信令。在多数情形中，带内信令意味着控制和管理流量是在用户流量被发送的同一信道上中继的，而带外信令意味着控制和用户流量是在不同网络上发送的，就像多数 PSTN 使用基于 7 号信令系统（SS7）的信令网络的情形一样。

由路由器交换的信息用于计算路由表和转发表。虽然或多或少的，路由表和转发表指代用户流量的转发路径，就路由器实现而言，在这两者间存在相当的区别，正常情况下，无论何时在网络的状态中存在一次变化，例如由于链路失效或节点失效导致的网络拓扑的一次变化。用于用户流量转发的路径可以是逐跳形式（从一个节点到下一个节点）路径或源路由显式形式的（从源到目的节点指定的完整路径）路径。

继续沿相同思路讨论，看看当网络拓扑中存在一次变化时，所发生的事件序列：

1）节点感知到一个或多个邻接节点或链路宕机。

2）之后该节点准备一条更新消息，广播或组播到邻接节点。

3）所有接收节点在接收到更新之后，重新计算路由路径。

4）一旦路径计算完成，则更新路由表和转发表。

5）路由更新被发送到其他节点。

如前面了解的，正在讨论分布式算法，所以时间成为最重要因素之一，它影响算法和路由协议的行为。上述的事件 1 到事件 4 涉及时间。在这些事件之间，总是存在一个短暂时间，期间就网络状态而言存在于所有路由器中的信息是不确定的，可能是不一致的。这可能导致路由协议行为不当，并在运作节点导致各种问题。如在任意分布式算法中，为了确保行为一致性，各定时器在路由协议中扮演一个重要角色。

如在前面讨论的，在确定路由协议类型及其行为中，网络拓扑扮演一个重要角色。当网络为静态的且在一个长时段上没有变化太频繁时，在路由表中可使用静态路由，这些路由可指定一次，并在必要时才人工改变。在网络状态（谱系）的另一端，情形是网络拓扑频繁地变化，则倾向于动态路由，其中当网络拓扑变化时，

路由表频繁地更新。

也存在像 BGP 一样的路由协议，它是基于路径矢量路由概念的。它类似于 DV，例外是它从每个节点接收距离度量指标和到目的地的整条路径。这有助于各节点检测路径中的任何环路。

将分组转发到下一跳涉及两项功能，即路由路径计算和分组交换。

路由路径计算涉及检查到一个目的主机或网络的所有可能路径，并寻找最优路由。在一个路由表中维护计算得到的路由路径。路由表中的信息包含细节，包括目的网络、下一跳和与路径相关联的一个度量指标。

分组交换过程涉及改变物理目的网络到下一跳（同时源和逻辑目的地址保持不变）。为使路由器确定如何路由一条分组，它需要如下信息：

1）目的地址。

2）邻居路由器。

3）到所有远端网络的可能路由。

4）到每个网络的最佳路由。

5）维护和验证路由信息的过程。

路由表中的路由可以是静态的或动态的。

静态路由用于固定拓扑网络，或管理员知道网络拓扑不会太频繁变化的场合。动态路由是通过在周期间隔接收到的更新从邻接路由器处学习得到的。静态路由过程的优势是在路由器处理器上路由计算没有开销，链路利用率降低，是安全的，原因是仅有管理员可改变配置。但是，这对于一个大型的、复杂网络（一直在变化）是不可扩展的，原因是任何时候只要有一个新节点要添加，则所有路由都必须人工地改变。

在路由表中，也定义了一条默认路由，它指向一台路由器，所有分组都要转发到该路由器，在路由表中没有定义针对它的显式路径。

在动态路由中，邻接路由器交换通过它们可达的远端网络的信息。之后接到信息的路由器计算最优路由路径，并更新路由表。当在网络拓扑中发生变化时，路由器将有关拓扑改变的更新发送到其他路由器。通过重新计算路径并通过路由更新重新分发这些路径，各路由器开始收敛。这些更新传遍整个网络。直到所有路由都收敛之前，重复这个过程。但是，在这种情形中，路由器需要消耗 CPU 处理能力和链路带宽，交换路由更新。为这个目的，定义了各种路由协议。

如在前面看到的，因特网以大概是层次结构配置的形式由相互连接的各 AS 组成。路由协议主要有两种，一种主要用在一个 AS 或域内，称作 IGP；一种用在域或 AS 之间，称作 EGP。

所有路由协议都定义通过各路由而路由分组要求多少时间或成本的一种度量。沿路径的每台路由器或节点被称作一跳。度量可以是路由所需的跳数，或它可以带宽或任何其他策略度量来加以定义。度量背后的思路是找到最佳路径。

网络中的各种路由器可运行不同路由协议，即使单台路由器有时也运行多个路由协议。

每个路由协议有一个独立的度量指标结构和算法，并可能与其他协议是不兼容的。在多台路由器正运行多种路由协议的一个网络中，路由信息的交换和选择最佳路径的能力是极端重要的。出于这个目的，当两台或多台路由器通过不同的路由协议通告相同的路径时，使用AD的概念确定一条路径的可靠性。AD是路由信息源的可信性的一个度量。AD仅有局部意义，不在路由更新中通告。AD值越小，则协议的可靠性就越大。

各种路由协议都有预先定义的默认AD。表3.6列出了思科支持的各协议的AD默认值。

表3.6　默认距离值表

路　由　源	默认距离值
连接的接口	0
EIGRP 汇总路由	5
外部 BGP	20
内部 EIGRP	90
IGRP	100
ODR	160
外部 EIGRP	170
内部 BGP	200
未知	255

注：ODR—应需路由。

如果AD是255，则路由器不相信那条路由的源，在路由表中不安装该路由。路由协议可被分为三个不同类型：

1）DV：使用距离，正常情况下是以所需跳数度量的，来寻找最佳路由。各路由器寻找最少跳数来确定最佳路径。这些协议中的两个协议是RIP和IGRP。

2）链路状态：基于Dijkstra的SPF算法。这个协议使用三个不同的表：邻接路由器的邻接关系、跟踪整体网络拓扑的一个链路状态数据库（LSDB）和一个路由表。这种类型协议的主要例子是OSPF。每台路由器将其接口的状态发送给网络中的每台其他路由器。在一次网络拓扑变化之后，链路状态路由协议非常快速地收敛，但对于路径计算却要求更多的带宽和计算能力。每次在网络拓扑中出现一次变化时，就触发更新。路由器对其数据库运行SPF算法，并产生网络的一个SPF树，自己作为树根。如果网络的一部分在另一部分之前接收路由信息，则收敛可能花费较长时间或SPF树和路由表会存储不准确的信息。更新包含时间戳和序列号。

3）混合法：使用 DV 和链路状态方法的一种组合法。这种协议的主要例子是 EIGRP。

DV 和链路状态协议之间的主要区别如表 3.7 所示。

表 3.7　DV 和链路状态算法的比较

DV	链路状态
如在邻接节点中所描述的，每个节点理解（整个）网络拓扑	每个节点从自己的角度理解网络拓扑
距离度量指标是沿路跳数的一次求和	节点从自己的角度计算到目的地的最短路径

交换和路由

在层 3 的互联是由路由器实施的。许多时候，路由器和层 3 交换机是同义使用的术语，但在现实中，这两种类型的设备之间存在一定程度的差异。

也存在各种模块或线卡，具有不同功能，插入交换机或路由器，实施防火墙、入侵检测和分组嗅探的功能以及性能分析。

一台层 3 交换机可实施一台路由器的多数功能。一台路由器的主要特征之一是能够转换从一种类型的网络到另一种类型的网络流量，并进行转发，如从 ATM 到以太网或从一个令牌环网络到以太网络。另一项区别性特征是层 3 IP 组播。一些交换机实现因特网组管理协议（IGMP）探测，以辅助层 3 的 IP 组播，并防止将组播流量广播到这样的端口，即没有设备注册那个特定的组播组流量。在一台层 3 交换机和一台路由器之间的主要差异是分组转发的物理实现。在路由之前，路由器有时使用深度分组检查来实施分组检查，而交换机仅能进行基于硬件的帧转发。一些功能包括基于逻辑 IP 目的地址的路径计算、在整条分组上实施校验和、处理分组中的任何选项信息、为会话提供有状态的安全和处理控制平面中的网络管理相关的功能。即使 QoS 相关的建议也可用于流量优先级划分，如在视频会议应用中的情形。

在较高级交换机中，如在层 4 及以上，术语“交换机”的含义是依据厂商而不同的（厂商特定的）。多数时间，这些包括诸如网络地址转换（NAT）和负载分配（依据 TCP 会话）等功能。这样的设备包括防火墙或因特网协议安全（IPsec）或虚拟专网（VPN）网关。一些应用，如扩展的访问列表，在端口号的基础上过滤分组，还有思科的 netflow 应用收集统计信息和计费信息（用于思科高端路由器）。

在层 7，交换机可指代这样的一台设备，如一个内容分发网络（CDN）中的一个 web 缓存。这些设备将实施如下功能，诸如在统一资源定位符（URL）基础上进行内容过滤。

3.4　主要路由协议

3.4.1　路由信息协议

RIP 是在 IP 网络中最早部署的路由协议之一。它基于 DV 算法或 Bellman-Ford 算法。它最初部署在 ARPANET 中，这是因特网的先驱网络。它是作为一个网关信息协议部署的，后来当它成为 Xerox 网络系统（XNS）（由 XEROX PARC 开发的）组成部分时，被重命名为 RIP。它后来被包括在伯克利软件分发版（BSD）UNIX 分发版中，作为 routed 守护进程。

RIP，作为一种 DV 协议，为计算路由信息，依赖于一种距离度量指标。为避免分组路由进入一个循环，将跳计数限制为 15。一条分组每次到达下一个节点时，分组中的跳计数加 1。当它达到最大跳计数时，分组被简单地丢弃。在 RIP 中，跳计数的 16 计数被称作无穷。但是，有一个有限的跳计数也影响 RIP 可在其中部署的网络大小。

RIP 分组（见表 3.8 和表 3.9）使用 UDP 发送，净荷类型设置为 520。相比诸如 OSPF 或 IS-IS 等其他协议，在快速拓扑变化的情形中，RIP 遇到慢收敛问题，并具有不佳的扩展性。通过调节跳限制 16（见图 3.10），开发了 RIP 的一个变种，这允许 RIP 部署在大型网络之中。在网络中 RIP 是相当容易配置的。

表 3.8　RIP 分组类型

分组类型	描　述
通用请求（操作类型 1）	在一台路由器出现在因特网之后，由之广播这条请求，学习网络上的所有节点。网络信息被设置为 0xFFFFFFF
特定请求（操作类型 1）	由路由器使用，得到有关一个特定网络的信息，目的是将分组路由给这个网络
周期性广播（操作类型 2）	确保所有节点都将网络拓扑信息保持最新。路由器也维护一个衰老（Aging）计数器，如果在某个时段上没有收到路由的任何更新，则清除这条路由
响应（操作类型 2）	作为对其他路由器的一条通用或特定请求的响应进行发送
特定信息响应（操作类型 2）	在网络上广播一个新服务的添加或去除

表 3.9　RIPv1 和 RIPv2 的比较

特　征	RIPv1	RIPv2
路由算法	DV	DV
路由更新	每 30s 进行周期性更新	周期性更新和改变时更新

（续）

特 征	RIPv1	RIPv2
广播/组播	广播到 255.255.255.255（MAC FF-FF-FF-FF-FF-FF）	组播到 224.0.0.9（MAC 01-00-5E-00-00-09）
度量指标	跳计数	跳计数
负载均衡	否	是
VLSM 支持	否	是
认证	否	是
限制	最大15跳计数（扩展性问题）	扩展性

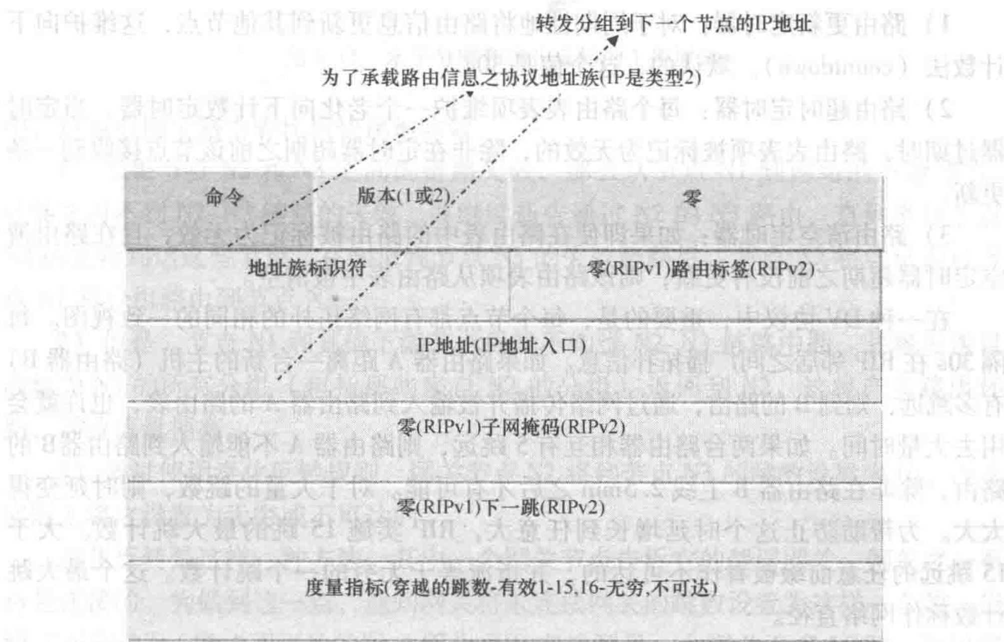


图 3.10 RIPv2：分组和协议

正常情况下，一台 RIP 路由器每隔 30s 传输一次 RIP 更新。但是，对于较大型的网络尺寸，以及路由表尺寸的增加，这将导致每 30 多秒的一次流量突发。在多数当前网络中，除非网络尺寸真正较小，否则并不真正使用或部署 RIP。

最初，每台 RIP 路由器每隔 30s 传输一次全更新。在早期部署中，路由表是足够小的，从而流量并不显著。但是，随着网络尺寸的增加，变得明显的是，每隔 30s 可能出现大量的流量突发，即使各路由器以随机时间初始化时也如此。人们过去认为，作为随机初始化的一个结果，路由更新将沿时间分散开来，但在实践中却不是这样的。

RIP 在 30s 的间隔上发送路由更新（默认的），同时当网络拓扑中出现一次变

化时也发送。当一台路由器接收到针对路由表中一个表项改变的一次更新时，它将新表项中的度量指标与老表项中的度量指标进行比较。如果新的度量指标小于较老表项的，则较老表项为新表项所替换。

本质上而言，所有网络协议是运行在多个处理器上的并行的、分布式算法。结果，在不同节点处的任务协作中，时间扮演一个重要角色。所以每种协议都维护多个定时器，确保各协议以一种稳定的和收敛的方式工作，即使在拓扑变化期间也是如此。一个稳定的和收敛的行为，意味着各节点在一个有限的时间内更新它们对网络拓扑的理解，且不影响网络性能。

RIP 也维护多个定时器。

由 RIP 维护的主要定时器有：

1）路由更新定时器：对于周期性地将路由信息更新到其他节点，这维护向下计数法（countdown）。默认的，这个值是 30s。

2）路由超时定时器：每个路由表表项维护一个老化向下计数定时器。当定时器过期时，路由表表项被标记为无效的，除非在定时器超期之前该节点接收到一条更新。

3）路由清空定时器：如果即使在路由表中的路由被标记为无效，且在路由清空定时器超期之前没有更新，则该路由表项从路由表中被清空。

在一种 DV 协议中，重要的是，每个节点都有网络拓扑的相同的一致视图。每隔 30s 在 RIP 邻居之间广播拓扑信息。如果路由器 A 距离一台新的主机（路由器 B）有多跳远，则到 B 的路由，通过网络传播并被输入到路由器 A 的路由表，也许就会用去大量时间。如果两台路由器相互有 5 跳远，则路由器 A 不能输入到路由器B 的路由，除非在路由器 B 上线 2.5min 之后才有可能。对于大量的跳数，则时延变得太大。为帮助防止这个时延增长到任意大，RIP 实施 15 跳的最大跳计数。大于 15 跳远的任意前缀被看作不可达的，并指派等于无穷的一个跳计数。这个最大跳计数称作网络直径。

因为像 RIP 的 DV 协议其工作方法，是周期性地将整个路由表洪泛到网络，则导致相当大的流量。像水平分割（Split Horizon）和毒化反转（Poison Reverse）等技术，可帮助降低 RIP 主机发出的网络流量总量，并使路由信息的传输更加高效。在防止路由环路中这也是有用的，这是当中间链路之一中断时出现的情况。

水平分割是在一个网络中防止路由环路的一种方法。基本原理是简单的：有关一个特定分组路由的信息永远不要发回到它被接收的方向。基本上说，这意味着，如果一台邻接路由器将一条路由发送到一台路由器，则接收路由器不要将这条路由在相同接口上发回通告路由器。称作水平分割的这项技术，帮助限制 RIP 路由流量总量，方法是去除在那个接口上其他邻居已经学习到的信息（见图 3.11）。

水平分割：当节点 N1 通过节点 N2 学习到节点 N3 的最短路由时，它不会将这条路由反向通告给节点 N2。这就防止了路由环路。在拓扑中出现一次变化的情形

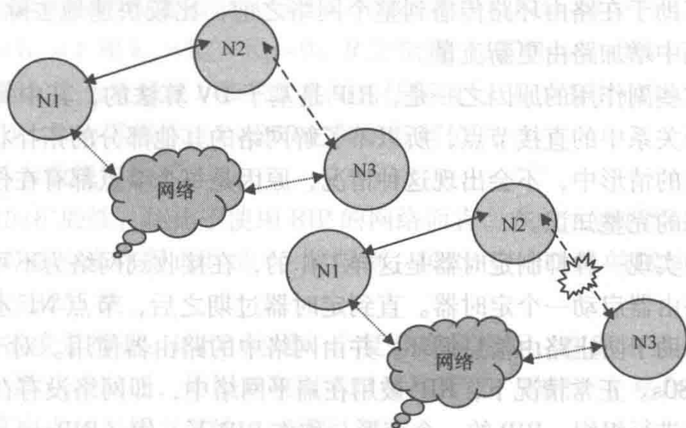


图3.11　水平分割和毒化反转的工作原理

中，这也有助于路由表比较快速地收敛。

1）如果节点N2和N3之间的链路失效，那么从节点N1到网络的下游节点，可能学习不到N2-N3链路的失效，并继续通告通过N2的N3路由，直到来自节点N2的更新到达这些节点。在知道到节点N3的失效链路后，节点N2将尝试通过节点N1将分组路由到节点N3。

2）但是，节点N1和其他下游节点仍然不知道N2-N3链路中断，并将发送目的地为N3的所有分组（包括那些来自N2的分组）返回到N2。这将产生路由环路，导致流量拥塞。

3）通过使用毒化反转规则，网关节点N2将到节点N3的跳数设置为16，由此实际上将之设置为无穷或不可达的。

毒化反转是这样一种方法，其中一个网关节点告诉它的邻居网关，网关之一不再是连接的。为做到这一点，通知网关将未连接网关的跳数设置为这样一个数，指明“无穷的”（指“不可达的”）。因为RIP支持到另一台网关至多15跳，所以将跳计数设置为16，将意味着“无穷的”。这是路由毒化所有可能反向路径的等价做法，即将针对一条特定分组回到源发节点的路径有一个无穷的度量指标这样的信息通知所有路由器。带有毒化反转的水平分割比简单的水平分割在有多条路由路径的网络中是更加有效的，虽然在仅有一条路由路径的网络中，这种做法没有提供比简单水平分割的任何改进。采用路由毒化，当一台路由器检测到连接的路由之一失效时，则该路由器将毒化该路由，方法是将一个无穷度量指标指派给它，并将之通告给各邻居。

当一台路由器向它的邻居通告一条毒化的路由时，它的邻居就打破水平分割的规则，并将相同的毒化路由发回源发者，称作一次毒化反转。为赋予该路由器足够的时间传播毒化的路由，并确保在传播时不发生路由环路，则各路由器实现一种抑

制机制。这有助于在路由环路传播到整个网络之前，比较快速地去除它们。但是，这也会在网络中增加路由更新流量。

注意，这些副作用的原因之一是，RIP 是基于 DV 算法的，其中每个节点简单地知道其邻居关系中的直接节点，所以不了解网络的其他部分的拓扑状态。在基于链路状态算法的情形中，不会出现这种情况，原因是每个节点都有在任何给定时刻网络拓扑状态的完整知识。

在 RIP 中实现一种抑制定时器是这样工作的，在接收到网络为不可达的信息之后，使每台路由器启动一个定时器。直到定时器过期之后，节点 N1 才丢弃任何路由更新。这有助于防止路由信息循环，并由网络中的路由器使用。对于 RIP，默认值被设置为 180s。正常情况下，RIP 被用在扁平网络中，即网络没有在任何区域以层次结构方式进行组织。RIP 的一个变形，称作 RIP 下一代（RIPng），可用于 IPv6 网络（RFC 2080）。

RIP 实现是 BSD 系统、GNU Zebra 和 Windows Server 以及思科路由器和多层交换机（运行 IOS 和 NX-OS）中 routed 的组成部分。

思科实现一个专有协议 IGRP，它类似于 RIP。这种情况也由 EIGRP 得以延续。

3.4.2 内部网关路由协议

IGRP 是一种高等 DV 路由协议。它是一个由思科开发的路由协议，仅有思科设备才将这个协议用于路由目的。

运行 IGRP 的一台路由器默认地每隔 90s 发送一次更新广播。当在三个更新时段内（270s）没有收到来自源发路由器的一次更新时，它就宣称一条路由无效。在 7 个更新时段（630s）之后，这包括 3 个更新时段，则路由器从路由表中去除该路由。IGRP 通告三种类型的路由：

1）内部：附接到一个路由器接口的网络中各子网之间的路由。

2）系统：到一个 AS 内各网络的路由。路由器从直接连接的网络接口和由其他使用 IGRP 通信（IGRP-speaking）的路由器提供的系统路由信息，推导得到系统路由。

3）外部：到 AS 外部各网络的路由，当识别无奈之选的一台网关时考虑这样的路由。路由器从 IGRP 提供的外部路由列表中选择无奈之选的一台网关。如果路由器没有一条分组的较佳路由且目的地不是一个连接的网络，则该路由器使用默认网关（路由器）。

IGRP 使用互联网络链路特征的一个复合度量指标，这些特征如时延、带宽、可靠性、MTU 和负载，确定每条分组的最佳路径。

使用如下公式选择最佳路径：

$$M=[(K_1\times B+K_2\times B)/(256-L)+K_3\times D]\times[K_5/(R+K_4)]$$

在比较简单的形式中，它可规定为 $M=(10^7/B)+D$。

其中，M 为一条路径的度量指标（最小的是较佳的）；$K_i(i=1\sim5)$ 为常数，正常情况下，$K_1=K_3=1$ 和 $K_2=K_4=K_5=0$；B 为带宽（1200bit/s ~ 10Gbit/s，相对 10^7 做归一化处理）；L 为负载（1 和 255 之间的任何值）；D 为拓扑时延（沿路径时延之和，$1\sim2^{24}$）；R 为可靠性（1 ~ 255 之间的任何值）。

其中的各参数规定如下：

1）增加的扩展性：相比于使用 RIP 的网络而言，在较大型网络中的路由有所改进，IGRP 可被用来克服 RIP 的 15 跳限制。IGRP 有 100 跳的默认最大跳计数，可配置为最大的 255 跳。

2）复杂的度量指标：IGRP 使用一个复合度量指标，在路由选择中提供相当大的灵活性。默认情况下，使用互联网络时延和带宽，得到一个复合度量指标。在度量指标计算中也可包括可靠性、负载和 MTU。

3）多路径支持：IGRP 可在一个网络源和目的地之间最多维护 6 条不等成本的路径，但仅有最小度量指标的那条路由被放置在路由表中。另外，RIP 仅维护具有最佳度量指标的那条路由，忽视其他路由。多路径可被用来增加可用带宽或路由器冗余（备份）。

4）不管拓扑如何变化均可做到快速收敛。

5）处理多种类型服务的能力。

IGRP 可用在这样的 IP 网络中，它要求一种简单的路由协议，但比 RIP 要鲁棒和具备扩展能力。同样 IGRP 可被配置执行触发式更新，由此降低了网络中路由更新流量。IGRP 也使用诸如水平分割、抑制定时器（280s）和毒化反转更新等技术来避免路由环路。

上面推导得到的复合度量指标被用来实施多条数据路径间的负载均衡。如果这两条路径具有复合度量指标数值 1 和 2，那么流量将在 2 到 1 比例的路由之间分配，即具有度量指标 1 的路由将携带度量指标为 2 的路由之流量的 2 倍。对于一个目的网络，IGRP 最多接受 4 条路径。这被称作不均匀负载均衡。思路是为最大化吞吐量和可靠性而分配流量。

像在 RIPv2 中一样，IGRP 也为拓扑变化期间改进稳定性和收敛而提供各种特征。这些包括：

1）水平分割。

2）瞬态更新：默认情况下，EIGRP 每 90s 发出一次更新。路由无效定时器被设置为 270s，而路由清空定时器被设置为 7 × 90s。当拓扑变化时，瞬态更新被用来比周期更新时段要早地发出路由更新，以加速收敛。

3）抑制定时器：在这个时段期间，放置路由，从而在某个时间段中路由器既不通告路由也不接受路由的通告。这防止在整个网络扩散不正确的或失步的更新。

4）毒化反转更新。

就默认路由而言，IGRP假定在AS边界上的路由器将有路由流量到AS外部的更完整信息，所以默认路由是到最佳边界路由器的路径。这个IGRP提供真实网络前缀的表项，正常情况下甚至是多个这样的边界路由器，而不是0.0.0.0作为哑默认网络，被指派为默认路由。这多个默认路由被周期性地扫描，以便选择具有最小复合度量指标的一条路由作为默认路由。对于默认路由器，默认的最大跳直径是100跳，而在IGRP中支持的最大距离是255跳。

3.4.3 增强的内部网关路由协议

不像看到的其他DV协议（即RIP和IGRP）的是，EIGRP不使用周期性的路由拓扑更新。仅当网络拓扑中存在变化时，才触发更新。在RIP和IGRP中，当一条路由丢失时，将其从路由表中清空，这取决于如下事实，即这些协议提供周期性更新。因为没有周期性更新，所以它利用一个hello协议，即发送周期性的hello分组，建立邻居关系，并检测一个邻居节点的丢失。存在EIGRP的两个主要版本0和1。基本上来说，这是一个高等DV协议。它也被称作一个混合协议，原因是它使用从链路状态协议得到的一些技术。

在EIGRP中，存在一个独立的hello子协议和一种可靠的更新机制。这使路由器能够构建当前网络拓扑的一个数据库，由此使一种周期性更新机制成为不必要的，同时可帮助防止环路。在这种情形中，一台路由器也存储它所有邻居的路由表，所以它可快速地找到一条替代路径。如果没有找到替代路径，那么它就查询邻接节点。直到找到一条路径之前，传播该查询。没有周期性更新，仅当任何路径的度量指标存在一次改变时，才发送部分更新。

在诸如RIP和IGRP的DV协议中，在发生一条链路丢失的事件中，存在形成路由环路的可能性。当有关一条路由的丢失信息没有到达网络中的所有路由器时（由于更新被丢弃或简单地由于所要求的时间延迟导致的），就形成路由环路。没有及时接收更新的各路由器，将通过它们的周期性广播，注入不存在的路由。EIGRP为邻居之间的所有更新使用可靠传输。如果邻居没有发送有关接收更新的一条确认，则重传更新（消息）。

hello子协议使用IP组播地址224.0.0.10，并映射到MAC地址01-00-5E-00-00-0A。hello协议的使用以及以触发式更新替换周期性更新的做法降低了协议的带宽需求。像在IGRP的情形中一样，EIGRP也使用一个复合度量指标。

RIP和IGRP使用各种技术（如水平分割、毒化反转和抑制定时器）来防止路由环路，而EIGRP则使用漫射（Diffusing）更新算法（DUAL）。除了最小成本路径外，DUAL维护到每个目的地的无环路路径的一个表。这有助于得到非常低的收敛时间。

准确地像IGRP度量指标一样，计算EIGRP复合度量指标，之后将其乘以256。由此，EIGRP复合度量指标的默认表达式是

度量指标 = (BandW + 时延) ×256

其中准确地像在IGRP中一样进行计算得到BandW和时延。参数BandW是这样计算的，从到目的地的所有外发接口中取最小带宽，并以10000000除以这个数（最小带宽），而时延是到目的网络的所有时延值之和（在数十微秒量级）。

EIGRP度量指标是IGRP度量指标的256倍。当一个网络同时运行IGRP和EIGRP（如在从IGRP到EIGRP迁移期间）时，这种方便的转换变得重要起来。

时延是一个累积性的值，通过将与路径中每个分段相关联的时延相加计算得到的。对于等成本路由，默认地激活负载均衡。对于替代路径，采用不相等的度量指标，可配置的方差支持不均等的负载均衡。但是，这也可能导致在端节点处分组交付的乱序，同时有组装这些分组所需的相应处理开销。

与IGRP一样，通过修改各参数，EIGRP可使用其度量指标中的负载和可靠性。

EIGRP的主要特征有：

1）快速收敛：使用DUAL，确保在路由计算期间不存在环路，做法是在一次拓扑变化之后，允许在计算中所涉及的所有路由器进行同时同步操作，这就降低了收敛时间。运行EIGRP的路由器存储所有邻接路由器的路由表，从而能够容易地找到替代路由。如果没有找到一条替代路由，那么它就向其邻居查询一条替代路由。直到找到一条替代路由之前，一直传播这些查询。

2）有限的带宽利用：因为EIGRP不实施周期性更新，仅当拓扑或度量指标变化时，才在组播地址224.0.0.10上触发更新并发送部分更新，所以带宽利用得到相当降低。这确保仅有要求这些更新的路由器才接收到更新。

3）路由汇聚和VLSMS支持：EIGRP支持VLSM，子网路由是由边界路由器汇总的。但是，它也可配置成在任何边界、任何接口处汇聚路由。

4）多协议支持：EIGRP支持IP和非IP协议。

5）扩展性：EIGRP可扩展到大型尺寸的网络，相比RIP的小于15跳，它可高达200跳以上。

1. EIGRP操作

该协议的主要操作组件如下：

1）邻居发现/恢复：当路由器上电，动态地学习到有关邻接路由器和附接网络的信息时，使用这个过程。路由器维护一个邻居表，其中存储邻居的IP地址和所附接的网络（类似于OSPF邻接表的结构）。当邻居发送一条hello分组时，它为预期为可达的和可操作的邻居通告一个持有时间。如果在持有定时器超期之前，没有接收到hello分组，那么它假定拓扑发生了变化。类似于OSPF（一种链路状态协议），EIGRP使用hello协议寻找和维持邻居的邻接关系。为了路由器与一个邻接路由器交换路由，它需要与之形成一个邻接关系。

2）可靠的传输协议（RTP）：所有EIGRP分组都是按序和可靠地交付的。

EIGRP也支持混合的单播和组播分组流。

3）采用组播的部分和增量触发式更新：在拓扑中任何变化之后，更新被立刻组播到仅受影响的路由器。并不发送完整的路由表，仅传输变化。将在下一节讨论组播路由。

4）DUAL有限状态机：这些算法用于降低计算时间的目标，并使所有受影响的节点同时计算新的路由路径。它跟踪由所有邻接节点通告的所有路由。DUAL选择一组可行的后继。一个后继是一个邻接节点，它用于以最低成本进行分组转发。确保这条路径是无环路的。当没有可行的后继时，假定网络拓扑发生了变化，则发生重新计算。DUAL尝试寻找一个可行的后继。如果通过使用路由表，为该节点和邻接节点找到一个可行的后继，则不需要进行重新计算。这有助于降低收敛时间。对于非广播多址（NBMA）网络，hello时间周期被设置为60s，而对高速NBMA网络，设置为5s。

2. EIGRP DUAL

给定作为DV协议组成部分提供的信息，可确定一条路径是否无环路。EIGRP分组是封装在IP中的。IP-EIGRP模块实施分组的封装以及剖析、发送和接收，并就接收到的任何新信息而通知DUAL。之后IP-EIGRP也重新发布由其他IP路由协议学习到的路径。EIGRP传输更新的速率取决于链路带宽。它支持每个接口为路由更新而使用的最大带宽百分比进行配置。即使在繁重流量时段，链路带宽的一个固定部分也可用于路由更新流量。

一个NBMA是这样一个网络，它支持多条路由器连接，但不支持广播或组播分组交付。这些网络有诸如帧中继或X.25网络等。邻居表表项的分组交付是使用一种可靠的传输机制发送的。这种机制使用序列编号、重传和往返定时器，确保各分组可靠地向邻居传输更新。

拓扑表由IP-EIGRP模块更新，由DUAL有限状态机使用。拓扑表包含到每个可达网络的距离和矢量集合所需的信息。一条特定路径的信息，由邻居报告组成，包括总时延、路径可靠性、路径MTU、可行距离（FD）、通告的距离（AD）和路由源。DUAL使用这个信息计算各后继和可行后继。

一个后继是这样一个邻居节点，在从拓扑表中计算可行路径之后，该节点被选作一个目的地节点的下一跳。一个可行后继是这样一个邻居，它满足可行性条件，并具有到目的节点的一条路径。一个可行性条件是要满足的一个条件：邻居的通告成本小于当前后继的成本。

当存在一个可行后继时，一条表项从拓扑表复制到路由表。从源到目的节点的多条路径形成FD的一个集合。所有邻居有小于FD的一个通告链路成本，那条路径被标记为无环路的。

AD指从后继到目的网络或节点的距离。当一个邻居改变它已经通告的度量指标时，除非在拓扑表中没有找到其他可行后继，否则它不会触发一次重新计算。带

有一个有效后继的一个节点被称作处于被动状态。拓扑表在 OSPF 中扮演邻接关系的角色，并向节点提供网络的一个准全局视图。如果没有找到一个可行后继，那么该协议触发路由更新和重新计算。路由重新计算以路由器向所有邻居发出一个查询开始。如果邻接路由器发现一个可行后继，那么就发回这个信息。称处在寻找一个可行后继过程的一个节点为处于主动模式。当处在主动状态时，该节点不能使用下一跳邻居转发分组。当接收到所有应答之后，找到一个可行后继或一组后继，那么该节点返回被动模式（见图 3. 12）。

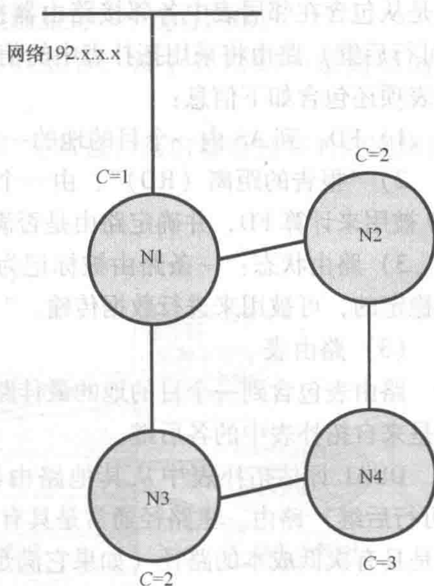


图 3. 12　EIGRP DUAL

如果 N1 和 N2 之间的链路失效，则节点 N2 需要另一条可行路由到达网络 192. x. x. x。其中节点 N2 将一条查询发往下一个邻接节点 N4。N4 注意到它有到 N1 的一个可行后继，由此就可达网络 192. x. x. x，所以它将这条可行路径发给 N2。如果 N2 没有一条可行路径，则它将不得不启动路由重新计算。

如果 N1 和 N3 之间的路径失效，那么除 N4 外 N3 就没有其他后继，因为通过 N4 的成本显示为 3，这高于 N3 到达网络 192. x. x. x 的当前成本。现在 N3 需要查询其后继，寻找一条可行路径。但是，N4 是唯一的邻居。当 N4 接收到该查询时，它不需要重新计算，因为它的邻居没有发生变化。所以它发出通过 N2 的可行路径，成本为 3。在 N3 接收到可行路径之后，因为 N3 的所有邻居现在已经重新计算路径，所以它安装通过 N4 的路径作为下一个可行后继，即使通过 N4 的成本大于通过 N1 的成本也这样做。

DUAL 使用 3 个独立的表进行路由计算。使用 EIGRP 路由器之间交换的信息创建这些表。这些信息不同于链路状态路由协议交换的那些信息。在 EIGRP 中，交换的信息包括路由、每条路由的“度量指标”或成本和形成一个邻居关系所需的信息（如 AS 号、定时器和 K 值）。这三个表及其功能的详细描述如下。

（1）邻居表

邻居表包含在所有其他直接连接的路由器上的信息。针对每个支持的协议（IP、IPX 等）存在一个独立的表。每个表项对应于一个邻居，带有一个网络接口和一个地址的描述。另外，初始化一个定时器，触发连接是否存活的周期性检测。这是通过 hello 分组做到的。如果在一个指定的时间时段期间没有从一个邻居接收到一条 hello 分组，则假定该路由器下线，并从邻居表中清除掉。

（2）拓扑表

拓扑表包含到AS内任何目的地的所有路由的度量指标（成本信息）。这种信息是从包含在邻居表中各邻接路由器接收到的。到一个目的地的主（后继）和辅（可行后继）路由将采用拓扑表中的信息进行判定。除各种信息外，拓扑表中的每个表项还包含如下信息：

1）FD：到AS内一个目的地的一条路由计算得到的度量指标。

2）"报告的距离（RD）"：由一个邻接路由器通告到一个目的地的度量指标。RD被用来计算FD，并确定路由是否满足"可行性条件"。

3）路由状态：一条路由被标记为"主动的"或"被动的"。"被动的"路由是稳定的，可被用来进行数据传输。"主动的"路由正在重新计算和/或不可用的。

（3）路由表

路由表包含到一个目的地的最佳路由（就最小"度量指标"而言的）。这些路由是来自拓扑表中的各后继。

DUAL评估拓扑表中从其他路由器接收到的数据，并计算主（后继）和辅（可行后继）路由。主路径通常是具有到达目的地最小度量指标的路径，而冗余路径是具有次低成本的路径（如果它满足可行性条件）。可能存在多个后继和多个可行后继。后继和可行后继都在拓扑表中维护，但仅有后继被添加到路由表，并用于路由分组。

要使一条路由成为一个可行后继，其RD必须小于后继的FD。如果满足这个可行性条件，则没有方法将这条路由添加到路由表（可能形成一个环路）。

如果到一个目的地的所有后继路由都失效，则可行后继成为后继，并被立刻添加到路由表。如果在拓扑表中没有可行后继，则发起一个查询过程寻找一条新路由。

思科EIGRP内的DUAL有限状态机蕴含着所有路由计算的判定过程。它跟踪由所有邻居通告的所有路由。称作度量指标的距离信息由DUAL用来选择高效的、无环路的路径。在可行后继的基础上，DUAL选择路由，插入到一个路由表。一个后继是用于分组转发的一个邻接路由器，它有到一个目的地的一条最低成本路径，确保不是一个路由环路的组成部分。当没有可行后继但有邻居通告目的地时，必须进行一个新后继的重新计算。它用来重新计算后继所用的时间量影响收敛时间。即使重新计算不是处理器密集的，但如无必要时避免重新计算，还是有好处的。当拓扑变化时，DUAL测试可行后继。如果存在一个可行后继，即避免重新计算。

1）DUAL维护到每个目的地的无环路路径的一个表。

2）DUAL将所有路径存储在拓扑表中。

3）它选择到每个目的地的最低成本的、无环路的路径。

4）DUAL支持EIGRP路由器确定由一个邻居通告的一条路径是有环路的还是无环路的，并支持运行EIGRP的一台路由器在不等待来自其他路由器更新的情况

下寻找替代路径。

5）EIGRP 采用4项关键技术，包括邻居发现/恢复、RTP、一个DUAL有限状态机和一种模块化架构，支持新协议容易地添加到一个现有网络。

6）一台 EIGRP 路由器从每个邻居接收通告，其中列出到一条路径的 AD 和FD。

7）AD 是从邻居到网络的度量指标。FD是从这台路由器通过邻居到网络的度量指标。

3. EIGRP：分组和协议

一个样例 EIGRP 分组结构如图 3. 13 所示。

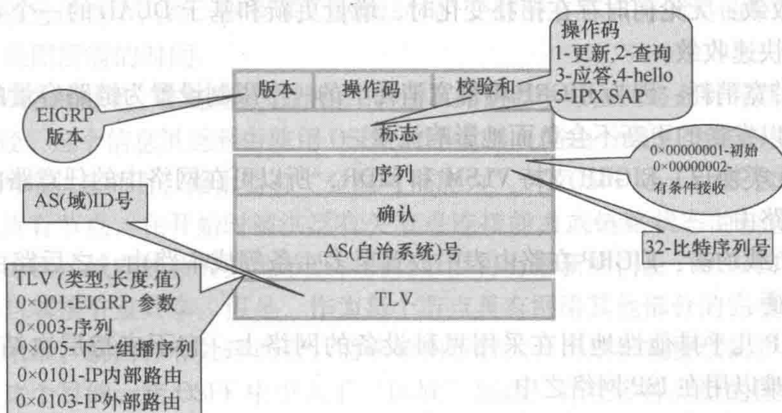


图 3. 13　EIGRP 分组格式

EIGRP 分组有 5 个类型：

1）Hello 分组：在 224. 0. 0. 10 上组播邻居发现/恢复，在快速链路上每隔 5s 发送一次，而在慢速和 NBMA 链路上每隔 60s 发送一次。像在 IGRP 的情形中一样，它也维护 hello 定时器和抑制（Hold）定时器。Hello 分组不要求确认。

2）确认：使用一个单播 IP 地址不可靠地发送，总是包含一个确认序列号。

3）更新：当由于一些拓扑变化而导致路由器从主动状态迁移到被动状态、存在一个目的节点的度量指标变化或当发现一个新节点时，发送更新。更新消息由路由器使用，构建它们的拓扑表。更新分组是作为单播分组发送的，是可靠地发送的。

4）查询：当路由器从被动状态变化到主动状态且该路由器被查询一个目的地的一条新的可行路径时，发送查询消息。查询可以级联（Cascade）通过网络，直到找到一条替代路径或到达一个网络边界。查询总是组播的，除非作为一条接收到的查询的响应，以一条级联的查询发送的。查询是可靠地传输的。

5）应答：在接收到一条查询之后，由每个 EIGRP 邻居发送的分组。应答是作

为单播分组发送的。

请求：为接收特定的信息，发送到各邻居。这些可以是单播的或组播的，并且是以不可靠的方式发送的。EIGRP 支持三种类型的路由：

1）内部路由：这些是从 EIGRP 学习到的路由。

2）外部路由：这些是从其他协议（如 OSPF 或 RIP）学习到的路由，可能来自其他 AS，之后在本地 AS 中重新发布。

3）汇总路由：这些是由 EIGRP 为自动汇总或作为路由汇总请求的响应，而动态地创建的。

4. EIGRP 的主要优势和劣势

1）收敛：无论何时存在拓扑变化时，增量更新和基于 DUAL 的一个查询过程都有助于快速收敛。

2）带宽消耗：因为 EIGRP 将带宽消耗上的一个限制设置为链路容量的一个百分比，所以发送的更新不会负面地影响流量。

3）无类协议：EIGRP 支持 VLSM 和 CIDR，所以可在网络中的任意路由器处过滤和汇总路由。

4）负载均衡：EIGRP 在路由表中放置至多 4 条等成本路由，之后路由器可实施负载均衡。

EIGRP 几乎排他性地用在采用思科设备的网络上。它不支持层次结构网络。结果，它难以用在 ISP 网络之中。

EIGRP 需要使用水平分割防止路由环路。当两台路由器首次成为邻居时，它们交换拓扑表。在这种情形中，由一台路由器接收的每条路由表项被通告回到邻居，带有最大度量指标值（毒化该路由），以便防止形成环路。当路由器从一个后继（用作查询中的目的地）接收到一条查询或一条更新，也可发生这种情况。

3.4.4　开放最短路径优先

1. 距离矢量和链路状态路由算法

迄今为止讨论了什么是 AS，并讨论一些路由协议（RIP、IGRP 等）。OSPF 是一个广泛使用的路由协议，由各种 AS 使用，在特定区域内或属于一个企业的多个区域中路由流量，并通过因特网连接。当一个企业的多个地点通过网络连接时，这可被看作单个 AS[11,12,18-21]。

OSPF 是基于从 Dijkstra 的 SPF 算法派生得到的链路状态算法。这里将简短地描述这个协议如何区别于 DV 协议。

DV 路由基于两个参数：距离或度量指标的概念（指定为到达一个目的地的跳数）和矢量（将分组向目的路由的方向，或接口/网络掩码）。在这种情形中，仅有直接连接的节点或路由器才共享距离和矢量信息。但是，距离大于 1 跳的各节点将不会共享这个信息，所以节点的“可视能力”是受限的，或仅具有“局部可视

能力”。这导致诸如路由环路的问题，原因是如果数跳远的某个节点被禁止，则该信息不会足够快速地传播到所有相关的节点。在这样一种情形中，需要使用治愈措施，诸如水平分割或毒化反转。

明显地，这些路由器需要共享一些信息，以确保任何网络状态（如节点断开或新节点加入）信息与其他节点周期性地共享。在DV协议的情形中，各节点周期性地将整个路由表发送到邻接节点。因为在这种情形中，各路由器不需要维护网络中所有链路的状态，路径或DV协议典型地要求较少的开销，这是就内存和处理而言的。但是，因为每个节点或路由器将仅有局部可视能力，由此就网络状态而言是相对有限的局部感知，所以网络状态（节点加入、去除）中的任何变化要用去更多时间传播通过整个网络，并影响“收敛时间”，即所有节点有网络的一个“共同共享的”视图所需的时间。

与DV路由相反，在链路状态路由中，所有路由器学习由网络上所有其他路由器可达的路径。这个信息洪泛到由使用OSPF的路由器（路由分组）共享的整个区域。因此，就网络状态而言，路由器有“全局可视能力”。在OSPF处于活跃状态的一个区域内的所有节点，在开始时被洪泛有关节点连接能力或链路状态的信息。但是，此后，假定就区域内的所有其他节点，所有节点将有最新的信息，所以仅有任何状态更新才与邻接节点共享。但是，作为每个节点具有网络其他部分的完整知识的这种范型的结果，就内存和处理而言，需求倾向于较大，所以影响协议的扩展性。

出于这个目的，在OSPF中引入了“区域”概念。存在多种类型的区域，并交换多种类型的分组，称作LSA，被包括作为OSPF规范的组成部分。

但是，因为每个节点或路由器都有全局可视能力，并因此共享网络状态的一个共同视图，所以网络状态（节点添加、去除）中的任何变化会更快速地传播通过网络，因为仅有更新过的信息才需要洪泛到其他的节点，这就降低了收敛时间。

一旦各节点具有一个区域中节点连接状态的所有信息，则每个节点就利用SPF算法，为在网络中分组转发计算相对于那个节点的最佳路径。

基本上而言，DV范型从“局部”节点收集信息，查看在每个节点处看起来的最佳路径是什么样子的，并使用它转发各分组。在链路状态范型中，思路是将网络或区域的详细“状态”传递到路由器/节点，并使它们计算得到转发分组的最佳路径。

2. 开放最短路径优先：操作

OSPF是一种IGP，利用链路状态算法来计算分组转发表。利用OSPF替换RIP或IGRP的主要原因之一是它可扩展到一个相当大型的网络，在任何拓扑变化之后收敛非常快速。

在深入研究OSPF的内部工作机理之前，将讨论一些基本术语。如所讨论的，一个AS的概念指一个站点或通过因特网链路绑定的一个企业的多个站点。但是，采用OSPF，随着网络尺寸增长，内存和计算处理开销也增加，所以有必要甚至将一个AS分解成多个区域。一个区域简单地是一个AS内路由器的一个层次结构，

它们形成邻居关系。

一台路由器由一个 router ID（路由器 ID）指明，这是指派给它的一个32比特的唯一数字。当两台路由器共享一条共同的链路时，称它们为邻居，而当两台路由器共享各 LSA 时，称它们为邻接关系。邻居未必形成邻接关系。在 OSPF 中，通过在网络中洪泛 LSA，各路由器交换信息，LSA 描述一条链路内的路由。它们使用来交换 LSA 的协议被称作 hello 协议。

OSPF 已经被指派为 IP 中的一种特殊协议类型，因为它工作在 IP 之上的层4。BGP 使用 TCP，与此不同的是，OSPF 不使用像 TCP 或 UDP 的任何其他传输层协议。OSPF 使用内建的确认和校验和，确保各分组没有丢失或是重复的。

OSPF 层次结构内的各路由器被分类并给定不同指示（Designation），这取决于它们所实施的功能。我们已经看到，OSPF 比多数 DV 协议有高得多的计算和内存需求。出于这个目的，一台相对功能强大的路由器保持子网的完整链路数据库并发送它的更新。这台路由器被规定为一台指定的路由器（DR）和被指定路由器（BDR）的一个备份。

也看到，在 OSPF 层次结构内存在多个区域，所以基于这些路由器的位置，在这些区域内的各路由器可扮演不同角色。位于区域0和一些其他区域的边界处的一台路由器被指定为一个区域边界路由器（ABR）。在两个或多个 AS 的边界处也存在一些路由器，它们的主要目的是将路由从一个 AS 重新发布到其自己的 AS，这些路由器被称作自治系统边界路由器（ASBR）。

将 AS 分成各区域基于各种因素，基于编址、区域尺寸、拓扑和/或策略。这些区域可基于一个特定区域内地址前缀或路由器数量（大约从50到300），从而不使之太大或太小，或物理连接的最小化，从而降低 ABR 的数量，甚至基于策略，即基于各种安全策略、组织等而分割（Segregating）流量。

将查看一个特定区域内的各 OSPF 路由器。当一台路由器启动后运行 OSPF 时，它开始发送 hello 分组，以便发现邻接路由器。这个过程的组成部分也涉及寻找一台 DR。每条 hello 分组包含链路状态和邻居的一个列表。正常情况下，功能比较强大的路由器应该是一台 DR 或一台 BDR。正常情况下，DR 的选举可以是“仓促形成的”，方法是指定一个较高的优先级，从而使期望的路由器成为 DR，或者具有较高 IP 地址的路由器获胜。DR 和 BDR 是负责为其所在区域产生一个 LSA 的那些路由器。其他路由器与 DR 和 BDR 交换 LSA，从而其他路由器不会因为相互发送 LSA 而过载，原因是对于 N 台邻接路由器，会终结于 $O(N^2)$ 问题。仅采用 DR/BDR 发出 LSA 更新，则变为一个线性 $O(N)$ 问题。LSDB 就像一个复制的分布式数据库，每台路由器维护它自己的数据库备份。在这样一种情形中，主要问题之一是保持数据库一致和在线性时间内保持同步。这是通过使用 DR/BDR 解决的问题。

同样在 OSPF 内有多个区域的目的之一是，限制在一个大型网络中采用 LSA 方式需要交换的信息总量。在一个大型网络中，由于 LSA 的周期性更新导致的流量，

可容易地成为流量的一大部分。

在一个区域内所有路由器的 LSDB 必须处于同步状态，否则将导致路由环路或黑洞。当路由器启动时，一台路由器的重要任务之一是同步 LSDB，方法是通过与其邻接和关联的路由器交换 LSA。

3. 区域的类型

一个企业的任何 OSPF 网络可被分隔成称为区域（Area）的子域（见图 3.14）。区域的类型见表 3.10。本质上而言，一个区域包含路由器和链路的逻辑连接，带有相同的区域标识，在相同区域内的各路由器维护对应于该区域的一个共同的拓扑。这些路由器没有该区域外网络拓扑的多少详细信息。这有助于将一个大型企业网络分成逻辑区域，并将各路由器维护的数据库尺寸降低。由此，OSPF 区域将一个层次式结构叠加在网络之上的数据流上。各区域被用来将路由器归组为可管理组，它们本地交换路由信息，但当向外部通告路由时汇总路由信息。

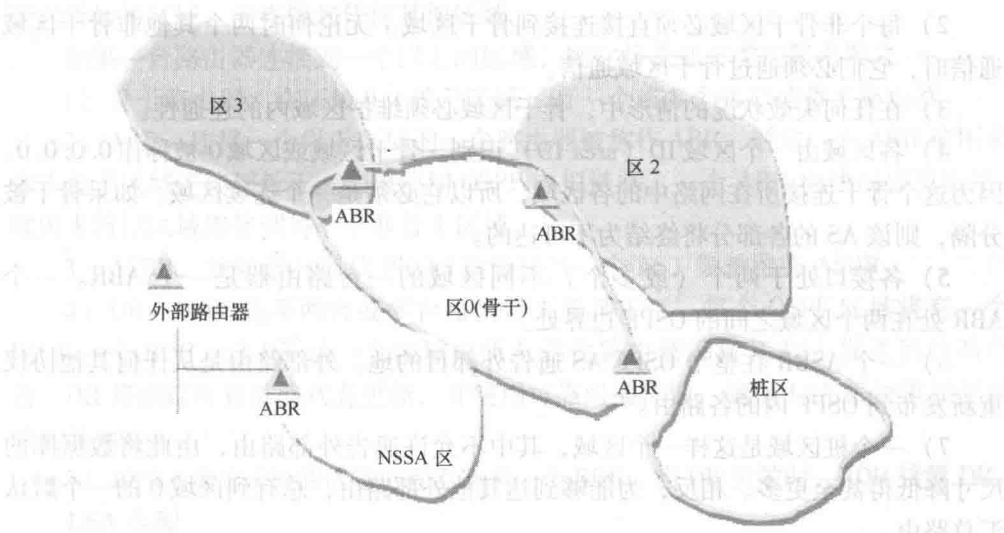


图 3.14 各种 OSPF 区域

表 3.10 区域的类型

区 域	限 制
正常	无
桩区（Stub）	不允许类型 5 AS 外部 LSA
纯桩区（Totally Stub）	除了默认汇总路由外，不允许类型 3、4 或 5 LSA
NSSA	不允许类型 5 AS 外部 LSA，但在 NSSA ABR 处转换为类型 5 的类型 7 LSA 能够穿越
NSSA 纯桩区（NSSA Totally Stub）	除了默认汇总路由外，不允许类型 3、4 或 5 LSA，但允许在 NSSA ABR 处转换为类型 5 的类型 7 LSA

- 每个区域有一个DR和一个BDR，这有助于在区域中洪泛LSA。通过汇总和过滤，不同区域之间的各路由器可交换路由信息。通过过滤和汇总路由，降低了要传播的路由数量。

将AS分成多个区域，有助于降低LSA的范围（类似于为降低以太网帧冲突的网桥域的某种概念）。被发送的一个LSA在这个区域洪泛，但取决于LSA类型，它不需要跨越到其他类型的区域。这也有助于降低OSPF LSDB的尺寸。

- 每个OSPF网络被分成如下描述的不同区域：

1）骨干是在使用OSPF的任何网络中应该总是构建的第一个区域，骨干总是区域0（零）。所有区域被直接连接到OSPF骨干区域。当设计一个OSPF骨干区域时，不应该存在由于一台路由器或链路失效而导致骨干区域被分成两个或多个部分的可能性。如果由于硬件失效或访问列表而导致OSPF骨干被分割，则网络的相当多的区域将是不可达的。区域0总是骨干或中心（Hub）区域。

2）每个非骨干区域必须直接连接到骨干区域。无论何时两个其他非骨干区域通信时，它们必须通过骨干区域通信。

3）在任何失效状况的情形中，骨干区域必须维护区域内的连通性。

4）各区域由一个区域ID（area ID）识别。骨干区域或区域0被称作0.0.0.0。因为这个骨干连接所在网络中的各区域，所以它必须是一个连续区域。如果骨干被分隔，则该AS的各部分将终结为不可达的。

5）各接口处于两个（或多个）不同区域的一台路由器是一个ABR。一个ABR处在两个区域之间的OSPF边界处。

6）一个ASBR在整个OSPF AS通告外部目的地。外部路由是从任何其他协议重新发布到OSPF内的各路由。

7）一个桩区域是这样一个区域，其中不允许通告外部路由，由此将数据库的尺寸降低得甚至更多。相反，为能够到达其他外部路由，总有到区域0的一个默认汇总路由。

8）桩区域被屏蔽于外部路由，并有到区域0的一条默认路由。但是，这些区域接收属于相同OSPF域其他区域之网络的信息。

9）一个纯桩区域仅连接到骨干区域。一个纯桩式/纯桩区域不会通告它知道的路由。它不发送任何LSA。一个纯桩区域接收的唯一路由是来自一个外部区域的默认路由，该区域必须是骨干区域。这条默认路由允许纯桩区域与网络其他部分通信。

10）桩区域仅被连接到骨干区域。桩区域不从AS外部接收路由，但确实从AS内接收路由，即使路由来自另一个区域时也如此。

11）频繁发生的情况是，一个独立的网络被用于将内部企业网络连接到因特网。OSPF为将一个ASBR放置在一个非骨干区域做好了准备。在这种情形中，桩区域必须从OSPF AS外部学习到路由。为方便做到这一点，定义了一种新的LSA类型——类型7 LSA。类型7 LSA是由ASBR创建的，并通过桩区域的边界路由器

转发到骨干。这允许其他区域学习 OSPF 路由域外部的路由。非如此桩区域（NS-SA）比桩区域更灵活，其中一个 NSSA 可将外部路由输入到 OSPF 路由域，由此向不是 OSPF 路由域组成部分的小型路由域提供中转服务。

此外，存在通过一条虚拟链路（有点像通过一个中间非零区域的隧道法）连接的区域。正常情况下，假定所有区域流量都通过区域 0，但在一些情形中，如果区域被悬挂（Hanging Off）其他区域并独立于区域 0 有相当的距离（拓扑上），则这就是不可能的。这个区域被添加，以便能够处理一些现有网络，其中使流量通过骨干是不可能的。

3.4.5　路由器类型

OSPF 路由器可作为各种角色，这取决于它们位于哪里以及它们参与哪些区域。

内部路由器：一台内部路由器仅连接到一个 OSPF 区域。它的所有接口都连接到它所处的区域，不连接到任何其他区域。

如果一台路由器连接到一个以上的区域，则它将是如下类型路由器之一：

1）骨干路由器。在区域 0（骨干区域）有一个或多个接口的骨干路由器。

2）ABR，连接一个以上区域的一台路由器被称作 ABR。通常一个 ABR 被用来将非骨干区域连接到骨干。如果使用 OSPF 虚拟链路，一个 ABR 也使用虚拟链路，被用来将该区域连接到另一个非骨干区域。

3）ASBR，如果路由将 OSPF AS 连接到另一个 AS，则被称作 ASBR。

4）DR：OSPF 选举两台或多台路由器来管理 LSA。每个 OSPF 区域将有一个 DR 和一个 BDR。该 DR 是一个区域内所有其他路由器都将其 LSA 发送到的路由器。DR 将跟踪所有链路状态更新，并使用可靠组播传输，确保 LSA 都被洪泛到网络的其他部分。

5）BDR：确定 DR 的选举过程也选举一个 BDR。当 DR 失效时，BDR 接管 DR。

LSA 类型

由运行 OSPF 的路由器交换的 LSA 分组使路由器保持它们的 LSDB 处于同步，并将更新导入路由表和转发表。每次在 LSDB 中出现一次变化时，路由器不得不使用 SPF 重新计算路由表。

对理解 OSPF 将如何影响网络至关重要的是，认识到存在多个 LSA 类型。每隔数秒发送更新，这导致对 LSA 数据库的更新，并可能更新路由表。“新的” LSA 将导致单台路由器丢弃其路由表，并开始采用 SPF 重新计算。

在“建立邻接关系”阶段，使用 Hello 和数据库描述。OSPF 分组类型 3 是一条链路状态请求，类型 4 是一条链路状态更新。最后，类型 5 是一个链路状态 ACK。OSPF 被实现为层 4 协议，所以它直接位于 IP 之上。既不使用 TCP 也不使用 UDP，所以为实现可靠性，OSPF 有一个校验和及其自己的内建 ACK。为通过嗅探流量而排错，需要知道 OSPF 组播地址是 224.0.0.5，而 DR 使用 224.0.0.6 实现相

互间通信。

存在6种不同的LSA分组类型（见表3.11）。

表3.11　OSPF协议的不同LSA类型

LSA类型	LSA描述
类型1：路由器LSA	这个LSA被用来报告路由器的活跃接口、IP地址和邻接关系。它仅在路由器的区域内分发，使用的是可靠洪泛技术（为一个新的LSA，通过捎带确认加以实现）。之后接收路由器再次在所有其他接口上发出
类型2：网络LSA	这个LSA由DR发送，并包含所有附接路由器的一个列表。使用该LSA，以便所有路由器不必将这个信息广播到另一台路由器，由此导致流量中的$O(N^2)$量级增加
类型3：汇总LSA	这个LSA是由ABR发出的，用来将其区域的汇总路由发送到其他附接的区域。对于目的区域的每个前缀，产生一个不同的LSA。发送到其他区域的这种路由汇聚，降低了接收路由器的LSDB的尺寸
类型4：ASBR汇总LSA	这些类似于汇总LSA（类型3），例外是这些LSA用于汇总源于另一个AS来自其他ASBR的路由。在这种情形中，ASBR将该LSA转发到一个ABR，之后该ABR将路径注入区域
类型5：AS外部LSA	这些LSA由ASBR洪泛到其他AS。这些LSA被用来由一个外部源通过其他路由协议（BGP、RIP等）学习到的前缀
类型6：组播路由器的组汇总	这些LSA由OSPF的组播扩展使用

3.4.6　边界网关协议

在前面一节中讨论了因特网的结构和拓扑，并考察了层1、层2和层3 ISP和这些ISP作为AS的情况。BGP是因特网的路由协议。本节将简短地考察一下BGP。BGP是一种EGP。它维护IP网络前缀的一个表，指定AS间的网络可达性。它没有使用基于度量指标的方法来确定链路成本，但路由决策却是基于指定的路径、策略和规则集的。BGP也提供多穴连接，其中各ISP可在多个点连接它们的网络（见图3.15）。

图3.16所示为一个典型拓扑，其中将部署BGP。BGP主要用于互联各种自治区域，包括由各层1或层2 ISP组成的区域。

如前面所讨论的，在较高层处流量特征和网络拓扑特征会影响一个路由协议的性能。在不同层形成因特网的多个AS，与作为一个AS组成不同的一个用户端相比，对路由流量的特征方面具有不同需求。在这些AS内的路由是由各种IGP完成的，如OSPF、RIP和EIGRP，而这些AS之间的所有路由是由BGP实施的。每个AS有一个唯一的AS号，且这些AS使用BGP，将其内部网络路由通告到其他AS。BGP被称作一种路径矢量协议，原因是它通告到达特定目的地所需的路径。在OSPF中个体路由器计算到整个网络的可达性，与此相反，BGP简单地通过到达目

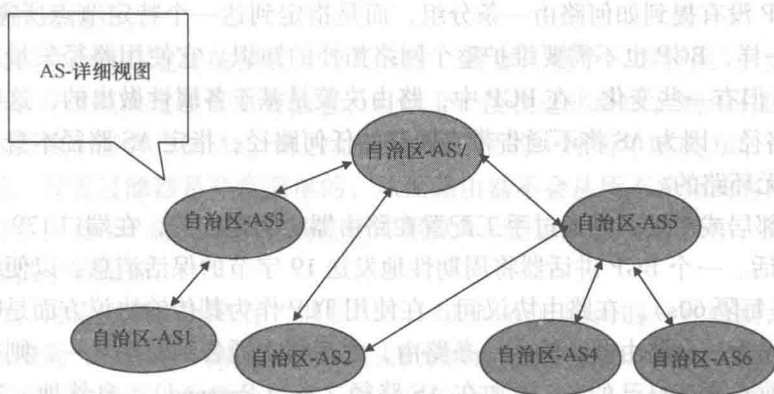


图 3.15　BGP 视图

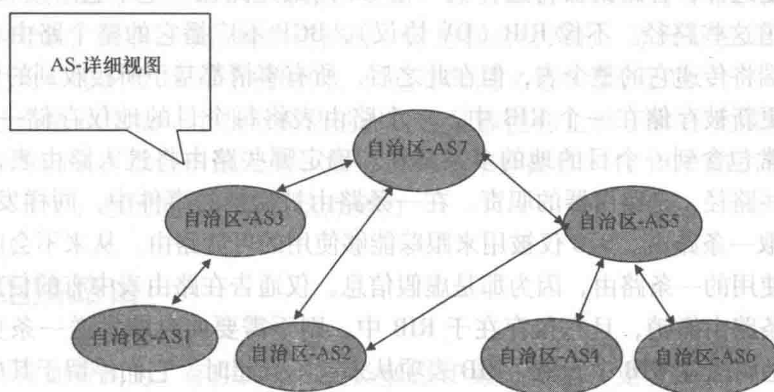


图 3.16　拓扑

的节点的路径的成本（以各种属性表示）实施计算。

实际上，BGP 是运行在 TCP 之上的，从最严格的意义上说，这使其成为一个 L4 协议。因为组成因特网的大型结构的层 1～3 AS 数量，相比整个因特网中个体节点的实际数量，是相对较少的，多数对端 AS 被直接连接到网络访问点（NAP）。同样，因为 TCP 处理多数面向连接的问题，所以就实现的复杂性和行为而言，相比 OSPF，BGP 作为一个协议是简单得多。在 BGP 中，两个对端需要维护一条连接，以便能够交换路由，且不对广播或组播做出应答。BGP 也没有任何发现机制。它取决于手工配置的路由。

BGP 的当前版本是 4，所以默认地总是谈论 BGP4。BGP 的较早期版本没有实施 CIDR 主机编址。

现在知道，因特网中的路由由两部分组成：内部细粒度的互联网络（由诸如 OSPF 的一种 IGP 所管理）和通过 BGP 的那些 AS 的互联。

因特网上的每个域至少有一个唯一的 AS 号，它们使用 BGP 将其网络通告到对

等端。BGP没有提到如何路由一条分组，而是指定到达一个特定节点所需的路径。像OSPF一样，BGP也不需要维护整个网络拓扑的知识。它使用路径矢量方法，类似于DV，但有一些变化。在BGP中，路由决策是基于各属性做出的，这些属性特定于AS路径。因为AS将不通告带有回路的任何路径，指定AS路径本身，确保这些路径是无环路的。

BGP邻居或对端，是通过手工配置在路由器之间建立的，在端口179上创建一个TCP会话。一个BGP讲话器将周期性地发送19字节的保活消息，以便维护连接（默认地是每隔60s）。在路由协议间，在使用TCP作为其传输协议方面是唯一的。

如果作为一台路由器，输入一条路由，之后将之通告到对端之一，则在通告这条路由之前必须将自己的AS添加在AS路径之前（Prepend）。自然地，这提供了人们可采用的一条"路径"，因为所通告的路由又离开了源AS一点。一般而言，但并不总是这样，各路由器将选择到一个AS的最短路径。基于它所接收的更新，BGP仅知道这些路径。不像RIP（DV协议），BGP不广播它的整个路由表。在启动时，对端将传递它的整个表，但在此之后，所有事情都基于所接收到的更新。

路由更新被存储在一个RIB中。一个路由表将每个目的地仅存储一条路由，但RIB通常包含到一个目的地的多条路径。确定哪些路由将进入路由表，即将实际使用哪些路径，是路由器的职责。在一条路由被撤销的事件中，同样发生的是，从RIB中取一条路由。RIB仅被用来跟踪能够使用的可能路由。从来不会向一个对端通告不使用的一条路由，因为那是虚假信息。仅通告在路由表中有的信息。如果接收到一条路由撤销，且它仅存在于RIB中，则不需要向对端发送一条更新；相反，静静地将之从RIB中删除。RIB表项从来就不会超时。它们停留于其中，直到我们认为那条路由不再有效时（才删除）。

在因特网上的大量路由可称之为是基于策略的。有时有一条价格高昂的链路，仅当必要时才想使用，或也许将有这样一条链路，可用之将流量仅发送到某些合作方。多数情况下，BGP属性"共同体"将被用来识别一个路由集合。如果希望让邻居知道有关一条路由的一些秘密信息，则在输出那些路由之前，可设置一个共同体号。这些号码是完全任意的，所以无论发送什么，都必须符合一个预先确定的要有某种意义。

另一个重要的BGP属性是多出口区分符（MED）。这被用来告诉一个远端AS，我们倾向于使用一个特定的出口点，即使可能有许多出口点时也如此。为得到BGP如何工作的一种真实感觉，重要的是花费一些时间讨论祸害（Plague）因特网的问题。

首先，对于路由表增长，有一个非常困难的问题。如果某人决定分解过去为单个/16网络的一个网络，他或她极可能要通告数百条新路由。当发生这种情况时，因特网上的每台路由器将得到每条新路由。人们会一直被迫使实施汇聚，或将多条路由组合成单条通告。汇聚并不总是可能的，特别当希望将一个/19分解成两个/20（将是地理上隔离的）时的情况。现在路由表正接近200000条路由，有时它们看

来会指数性地增长。

其次，总是存在这样的担忧，即某人将“通告（整个）因特网”。如果某个大型ISP的客户突然决定通告所有信息，且该ISP接收这些路由，则所有因特网流量将被发送到该小型客户的AS。存在处理这种情况的一种简单解决方案，它被称作路由过滤。设置过滤器是非常简单的，从而路由器不会从所不希望的客户处接受路由，但许多大型ISP将仍然从对端接受“默认”（路由）的等价操作，这些对端不太可能提供中转服务。

最后，讨论振荡。BGP有“抑制”（Hold Down）路由的一种机制，这些路由看来有点错乱（Flaky）。倒换或忽来忽去的路由，通常来说，要向之发送流量是不足够可靠的。如果路由频繁地振荡，则在所有因特网路由上的负载将增加，这是由于每次某个路由消失和重新出现时的更新处理导致的。抑制（Dampening）将防止BGP对端从振荡的对端侦听所有路由更新。处在抑制中的时间总量，随每次振荡而成指数增加。当有一条故障的链路时，这是有点恼人的，原因是在访问许多因特网站点之前可能花费1h以上的时间，但这个时间又是非常必要的。

这已经是BGP的一个非常简洁的讨论，足够使您正确了解有关该协议，但无论从哪个方面来说都是不深入的。如果您的任务是操作一台BGP路由器，花点时间阅读各RFC：您的同行将欣赏这点努力。

3.5 组播路由

如在前面看到的，存在四种不同类型的IP编址机制（见表3.12）。

表3.12 不同类型的编址和路由

单播流量	广播流量	组播流量	任意播流量
S → D	S → D1, D2, D3	S → D1, D2, D3	S → D1, D2, D3
单播IP地址指单个发送方和接收方。将相同数据发送到多个单播地址，要求多次发送相同数据	广播IP地址指这样的地址，被用来将相同数据发送到所有可能的目的地。这允许发送方向所有目的地仅发送数据的一份备份。同样，可完成一次受限广播，方法是将网络前缀与一个主机后缀均为全1组合使用，即对于一个192.0.2.x/24网络，受限广播地址是192.0.2.255	组播IP地址与一个接收方结合相关联。这些使用224.0.0.0到239.255.255.255范围中的地址。在这种情形中，L3层路由器实施如下操作，它复制分组，并发往所有接收方，这些接收方注册到一个特定的组播IP地址。组播被用于这样的应用，如高速视频、视频会议和股票更新，其中存在少量发送方和多得多的接收方	任意播IP地址也被用作一到多分组传输。但是，数据报被路由到网络中“最接近的”接收者。本质上这是一个IPv6概念，多数情况下用在DNS服务器中进行负载均衡

在组播中，涉及一个发送方和多个接收方或接收方群组。这个接收方群组被称作一个组播组（见表3.13）。在单播路由中，多数情况下，各分组仅通过路由器的接口之一发送，与此不同，在组播路由器中，路由器可通过多个接口转发外发分组（见图3.17）。在组播的情况下，使用多个单播，是不可行的。它是极端低效的，并在有大量客户端时，导致拥塞和时延。

表3.13 不同组播组

224.0.0.0	基 地 址	RFC 1112
224.0.0.1	在这个子网上的所有系统	RFC 1112
224.0.0.2	在这个子网上的所有路由	
224.0.0.4	DVMRP 路由器	RFC 1075
224.0.0.5	OSPFIGP 所有路由器	RFC 2328
224.0.0.6	OSPFIGP 各 DR	RFC 2328
224.0.0.9	RIP2 路由器	RFC 1723
224.0.0.10	IGRP 路由器	
224.0.0.12	DHCP 服务器/中继代理	RFC 1884
224.0.0.18	VRRP	RFC 3768
224.0.0.102	HSRP	

注：DVMRP—距离矢量组播路由协议；OSPFIGP—开放最短路径优先内部网关协议；DHCP—动态主机配置协议；VRRP—虚拟路由器冗余协议；HSRP—主机待机（Standby）路由器协议。

图3.17 单播和组播

组播用于一到多或多到多内容分发，主要用于多媒体内容，如调度安排的音频视频分发、推送媒体（新闻头条）、文件分发和缓存、宣言（Announcements）和股票价格监测（一到多风格的应用），以及多媒体会议，资源同步，如分布式数据库、协作式学习和开发、远程学习等（多到多应用）。为理解组播路由，需要熟悉如下概念：组播编址、组播组、组播路由树。

3.5.1 组播编址指派

地址范围 224. 0. 0. 0/4（从 224. 0. 0. 0 到 239. 255. 255. 255，高位比特是 1110）从来不被指派为单播地址。在当前情况下，多种策略被用于指派或分配组播组地址。地址范围 224. 0. 0. 0/24 仅被指派用于本地子网上的组播，寻址到这个目的地的分组从来不会转发到该子网外部。在子网中的各种其他地址被指派给不同应用。从 239. 0. 0. 0 到 239. 255. 255. 255 的地址范围被保留用于管理范围的地址，并在私有组播域中没有地址冲突条件下可由任何人使用（参见 RFC 2365 和 RFC 1918）。

每个地址空间、IP 地址和物理地址空间定义组播地址（见图 3. 18）。

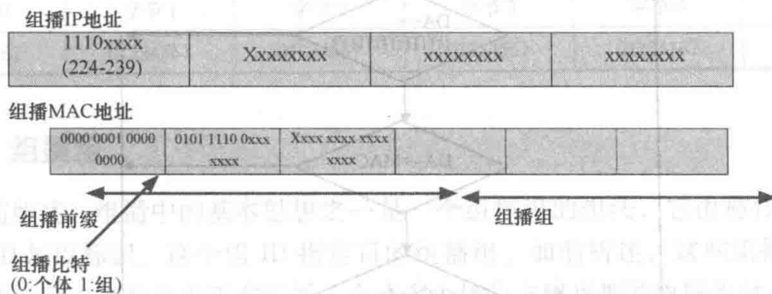


图 3. 18 组播地址

为组播分配组播块地址 01：00：5E：00：00：00 ~ 01：00：5E：7F：FF：FF。

所有 IP 地址被映射到某个以太网帧地址。单播分组被映射到特定的 MAC 地址。广播地址被映射到广播 L2 MAC 地址 FF：FF：FF：FF：FF：FF。正常情况下，IP 地址到以太网地址解析是通过地址解析协议（ARP）完成的，由于 IP 地址到以太网地址的唯一映射，这是可能的。但是，在组播 IP 地址的情形中，这出现了困难，这是由一个组播 IP 地址到多个以太网地址映射造成的。出于这个原因，在组播 IP 地址到以太网地址变换的情形中，不使用 ARP，而是直接映射到以太网 MAC 地址。

组播 IP 地址被映射到 01：00：5E：00：00：00 ~ 01：00：5E：7F：FF：FF 范围的以太网 MAC 地址（见图 3. 19），这仅为映射到组播 IP 地址留下 23 比特。因为一个组播 IP 地址的低 28 比特要被映射到这些以太网 MAC 地址的低 23 比特空间，这导致组播 IP 地址到以太网 MAC 地址之间的一些重复映射。由于此，如果存在订阅到多个组播组的主机，则网络层需要过滤掉来自其他组播组的分组，原因是这些主机没有订阅到这些组。

欲了解组播地址的一个完整列表（见表 3. 14），请参见 http：//www. iana. org/assignments/multicast- addresses。

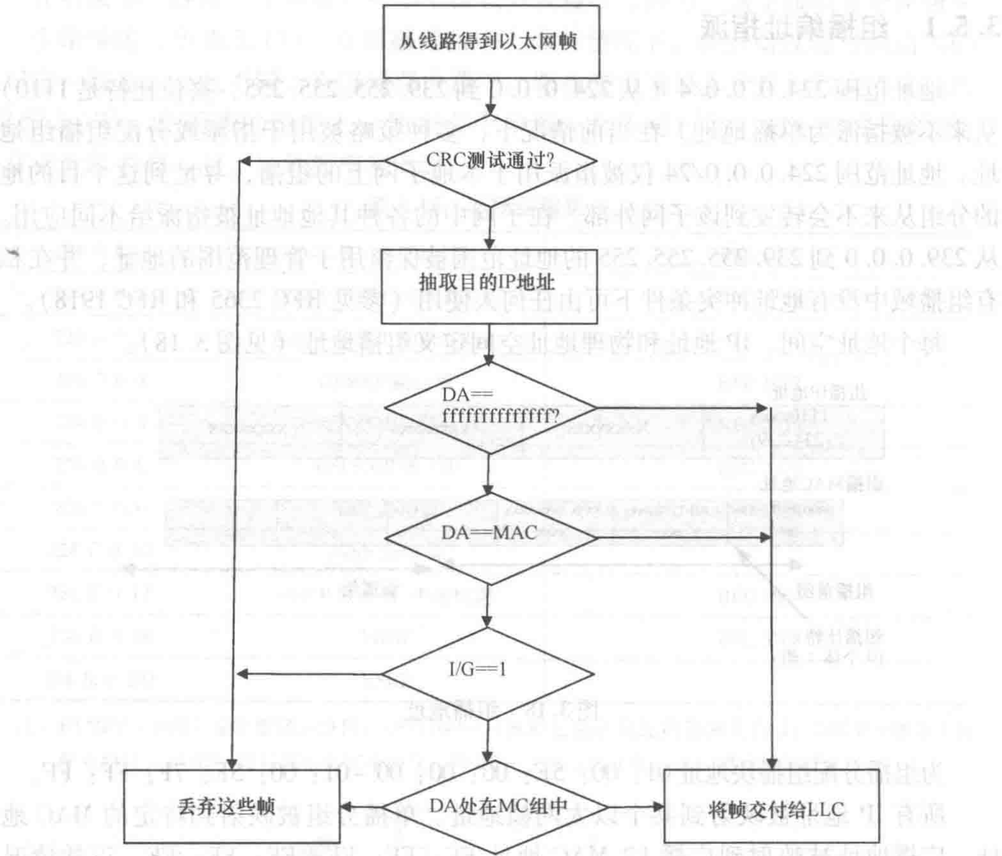


图3.19 将广播/组播IP地址映射到以太网MAC地址

表3.14 组播组

地　址	组	地　址	组
224.0.0.0	保留	224.0.1.7	AudioNews
224.0.0.1	在这个子网上的所有系统	224.0.1.10	IETF-1-LOW-AUDIO
224.0.0.2	在这个子网上的所有路由器	224.0.1.11	IETF-1-AUDIO
224.0.0.4	DVMP 路由器	224.0.1.12	IETF-1-VIDEO
224.0.0.5	OSPFIGP 所有路由器	224.0.1.13	IETF-2-LOW-AUDIO
224.0.0.6	OSPFIGP DR	224.0.1.14	IETF-2-AUDIO
224.0.0.7	ST 路由器	224.0.1.15	IETF-2-VIDEO
224.0.0.8	ST 路由器	224.0.1.16	MUSIC-SERVICE
224.0.0.9	RIP2 路由器	224.0.1.17	SEANET-TELEMETRY
224.0.0.10	IGRP 路由器	224.0.1.18	SEANET-IMAGE

注：可能有32个IP地址映射到同一个MAC地址。但是，冲突是极不可能发生的。

就目的地址中I/G比特的定位，经常出现混淆，原因是在内存中目的地址的比特顺序和在帧传输期间的比特顺序是不同的（见表3.15）。

表3.15　内存和帧传输中的组播地址

字节0	字节1	字节2	字节3	字节4	字节5
76543210	76543210	76543210	76543210	76543210	76543210

G/I字节顺序是预留比特顺序，在每个字节中被翻转（见表3.16）。

表3.16　G/I字节顺序

字节0	字节1	字节2	字节3	字节4	字节5
01234567	01234567	01234567	01234567	01234567	01234567

3.5.2　组播组

如前所述，组播中的基本思想之一是一个组播组的想法，它由称作组播组ID的一个ID加以标识。这个组ID指定目的组播组。如前所述，这些组播地址是D类地址的变形。当源节点正在组播一个或多个接收方感兴趣的数据流时，且如果这些接收方位于不同子网中时，那么这些接收方需要实施称作“加入组播组”的一项操作，这简单地意味着它们注册到中间路由器，这些路由器将组播分组转发到这样的主机，它们对接收与这个特定组播组ID关联的数据流感兴趣。当一台主机不再对数据流感兴趣时，它发送一条消息“离开”该组。这些功能是由因特网组管理协议（IGMP）实施的。

通过查看IGMP报告，中间路由器可判定在它们所连接的LAN分段上是否有任何接收方以及它们是否需要在这些网络分段上转发组播分组。

组的概念对组播的概念是至关重要的。由定义，一条组播消息是从一个源发送到目的地主机的一个组的。在IP组播中，组播组有称作组播组ID的一个ID。无论何时发出一条组播消息，则一个组播组ID指定目的组。这些组ID本质上是称作“D类”的一个IP地址集。因此，如果一台主机（一台主机中的一个进程）希望接收发送到一个特定组的一条组播消息，则需要以某种方式侦听发送到那个特定组的所有消息。如果一条组播分组的源和目的地共享一条共同的总线（即以太网总线），则每台主机仅需要知道在那台主机的进程间哪些组带有成员就够了。但是，如果源和目的地不在同一LAN上，则将组播消息转发到目的地就变得比较复杂。为解决组播消息因特网范围路由的问题，各主机需要加入一个组，方法是通知其子网上的组播路由器。IGMP用于这个目的。离开一个组也是通过IGMP完成的。采取这种方式，组播路由器可找出哪些主机是其网络分段上组播组的成员，并可确定是否在其网络上转发一条组播消息。

3.5.3　组播树

在组播路由中，必须满足如下这些需求：

1）每个组成员应该仅接收组播分组的一个备份，而不属于一个组播组的那些主机一定不能接收到任何备份。

2）一条组播分组一定不要一次以上地通过一台路由器。

3）从组播源到成员主机的路径必须是一条最优路径。

如在前面看到的，单播路由使用网络中的图（Graph）；在组播路由的情形中，使用树（见图 3.20）。这些树被称作生成树，具有最优路径的树被称作最短路径生成树。

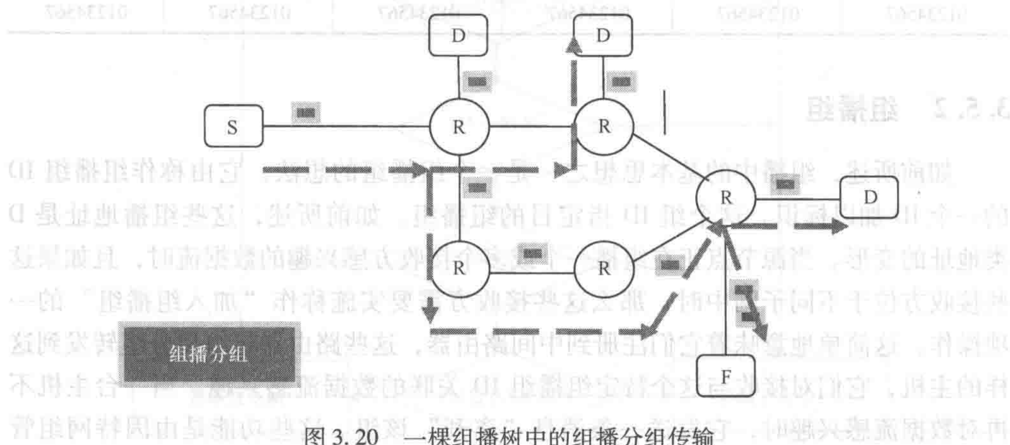


图 3.20　一棵组播树中的组播分组传输

到现在，读者已经认识到组播作为一种一个发送器到多个接收器的现象，分组分发是沿树发生的。支持组播的路由器创建分发树，它控制 IP 组播流量通过网络所取的路径，目标是将流量分发到所有接收方主机。存在两种类型的组播树：基于源的树和共享树。

要记住的另一个点是有关组播组成员协议的。网络中的各种主机必须让最近的路由器知道它们希望加入一个特定的组播组，或如果它们已经是组播组的成员时，它们要离开该组。这些协议被称作成员关系组协议，将在下面进一步讨论。

1. 基于源的树

一棵源树是一棵分发树的最简单形式。组播流量的源主机位于树根处，而接收方位于树枝的末端（见图 3.21）。组播流量从源主机沿树向下朝接收方传播。在哪个接口上应该发出一条组播分组的转发决策，依据的是组播转发表。这个表由一系列组播状态表项组成，它们被缓存在路由器中。一棵源树的状态表项使用表示 (S，G)。字母 S 表示源的 IP 地址，G 表示组播地址。

每个源有一棵树，由之产生一个组播组。一个源可有针对几个不同组的几棵

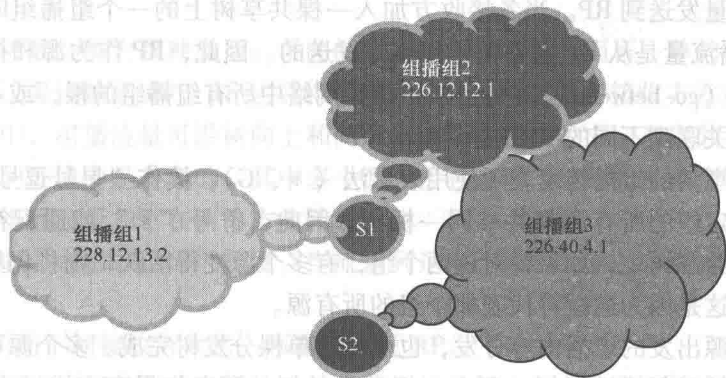


图 3.21　组播源树

树。如果存在 s 个源和 g 个组，那么不同的基于源的树，其最大数量为 $s \times g$。在基于源的组播树中，每个源有一棵单独的最短路径树，这支持以低的时延进行高效的分组路由。这也被称作密集区域组播（DM）。

下面是多棵生成树的一个例子，是针对不同组播组的，有两个不同的源（见图 3.22）。

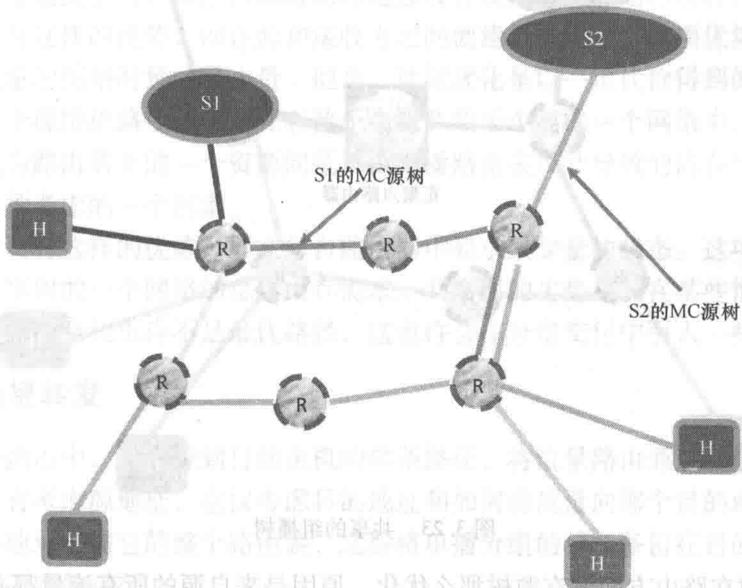


图 3.22　组播中的多棵生成树

2. 共享树

共享树区别于源树的是，树根是网络中某处的一个共同点。这个共同点被称作汇聚点（RP）。RP 是这样的点，各接收方在此处加入，学习到活跃的源。组播源

必须将其流量发送到RP。当各接收方加入一棵共享树上的一个组播组时，树根总是RP，组播流量是从RP向下朝各接收方发送的。因此，RP作为源和接收方之间的一个中介（go-between）。一个RP可以是网络中所有组播组的根，或不同范围的组播组可被关联到不同的RP。

一棵共享树的组播转发表项使用表示法（*，G），读作“星号逗号G”。这是因为一个特定组的所有源都共享同一棵树。因此，符号“*”或通配符代表所有源。在一棵共享树中，如果针对这两个组，有多个源变得活跃，则将仍然仅有两个路由表项，这是因为通配符代表那个组的所有源。

从多个源出发的组播内容分发，也可使用单棵分发树完成。多个源可共享单棵树。但是，针对每棵组播树，存在一棵独立的树。所以如果存在G个组，那么不管源的数量为多少，共享树的总数将是G。在这种情形中，每个源选择树中路由器之一作为一个RP（汇聚点）。如图3.23所示，该RP是源节点S1和S2的RP路由器。

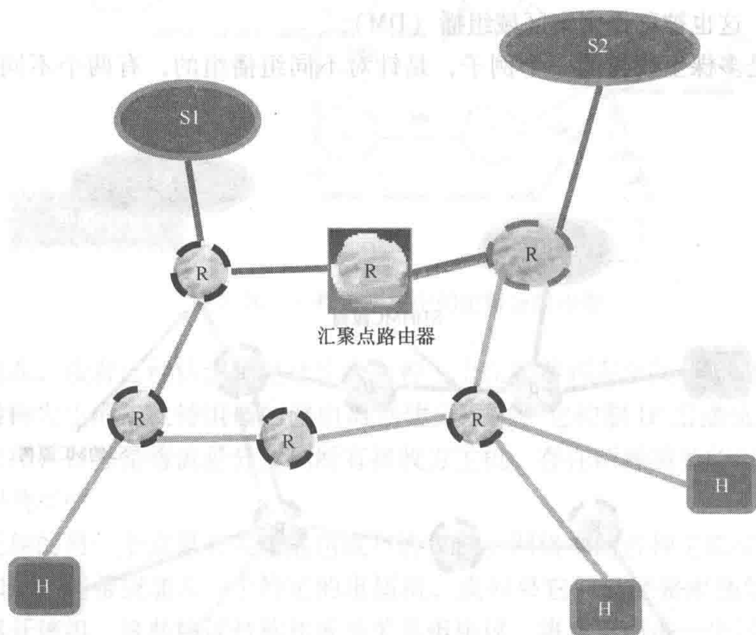


图3.23 共享的组播树

共享树在路由方面没有源树那么优化，原因是来自源的所有流量都必须传输到RP，之后遵循到各接收器的相同（*，G）路径。但是，所要求的组播路由状态信息总量小于一棵源树的信息总量。所以，在最优路由和必须保持的状态信息总量之间存在一项折中。

共享树支持接收端从一个组播组得到数据，而不必知道源的IP地址。需要知道的唯一IP地址是RP的IP地址。这可在每台路由器上静态地加以配置或动态地

学习。

共享树可被分类为两种类型：单向的和双向的。本质上而言，单向树是已经讨论的那种树：源传输到RP，RP之后将组播流量沿树向下朝各接收方发送。在一棵双向共享树中，组播流量可沿树向上和向下传输到达各接收方。双向共享树在一个任意到任意的环境中是有用的，其中多个源和接收方均匀地分布在整个网络中。来自源主机的组播流量如下在两个方向转发：

1）沿树向上朝根（RP）转发。当流量到达RP时，之后它沿树向下朝接收方发送。

2）沿树向下朝接收方发送。不需要通过RP。

双向树提供优于单向共享树的改进的路由优化能力，方法是能够在两个方向转发数据，同时保持最小的状态信息总量。

3. 源树与共享树

源树和共享树都是无环的。仅在树分支处才复制消息。

组播组的成员可在任何时间加入或离开，因此分发树必须加以动态更新。当在一个特定分支上不再有一个特定组播组的活跃接收方时，各路由器从分发树中剪掉那个分支，并停止沿那个分支转发流量。如果在那个分支上的一个接收方变得活跃，并请求组播流量时，该路由器将动态地修改分发树，并开始再次转发流量。

源树具有这样的优势，即在源和接收方之间创建最优路径。这项优势确保针对转发组播流量的网络时延的最小量。但是，这项优化是以一定代价得到的：各路由器必须为每个源维护路径信息。在有数千个源和数千个组的一个网络中，这项开销可快速地成为路由器上的一个资源问题。由组播路由表尺寸导致的内存消耗是网络设计人员必须考虑的一个因素。

共享树具有这样的优势，即在每台路由器中要求最少量的状态。这项优势降低了仅支持共享树的一个网络的总体内存需求。共享树的劣势是，在某些情况下，源和接收方之间的路径也许不是最优路径，这也许会在分组交付中引入一些延迟。

3.5.4　组播转发

在单播路由中，沿从源到目的主机的单条路径，将流量路由通过网络。一台单播路由器没有考虑源地址，它仅考虑目的地址和如何将流量向那个目的地转发。路由器以目的地址扫描它的整个路由表，之后将单播分组的单个备份在目的地的方向上转发出到正确的接口。

在组播转发中，源将流量发送到主机的一个任意组，该组由一个组播组地址表示。组播路由器必须确定哪个方向是上行方向（朝向源）和哪个方向是下行方向（或多个方向）。如果存在多个下行路径，则路由器复制分组，并将之向下沿合适的下行路径（最佳的单播路由度量指标）——未必是所有路径。将组播流量转发离开源，而不是到接收方，被称作反向路径转发（RPF）。在下一节描述RPF。

3.5.5 组播路由算法

1）存在正被开发和使用的各种组播路由算法。这些算法中的一些算法使用共享树，而其他算法使用基于源的路由树（RPF）。

2）反向路径广播（RPB）。

3）截断的反向路径广播（TRPB）。

4）反向路径组播（RPM）。

5）基于核的树（CBT）。

前四个算法使用基于源的组播路由，而最后一种算法使用共享树算法。

1. 反向路径转发

在这种情形中，单播路由表被用来沿从接收方到源的反向路径创建一棵分发树。之后组播路由器沿从源到接收方的分发树转发分组。这使路由器能够沿分发树向下正确地转发组播流量。RPF利用现有单播路由表，确定上行邻居和下行邻居。仅当一台路由器在上行接口上接收到一条组播分组时，它才转发该分组。这种RPF检查将有助于确保分发树是无环的。

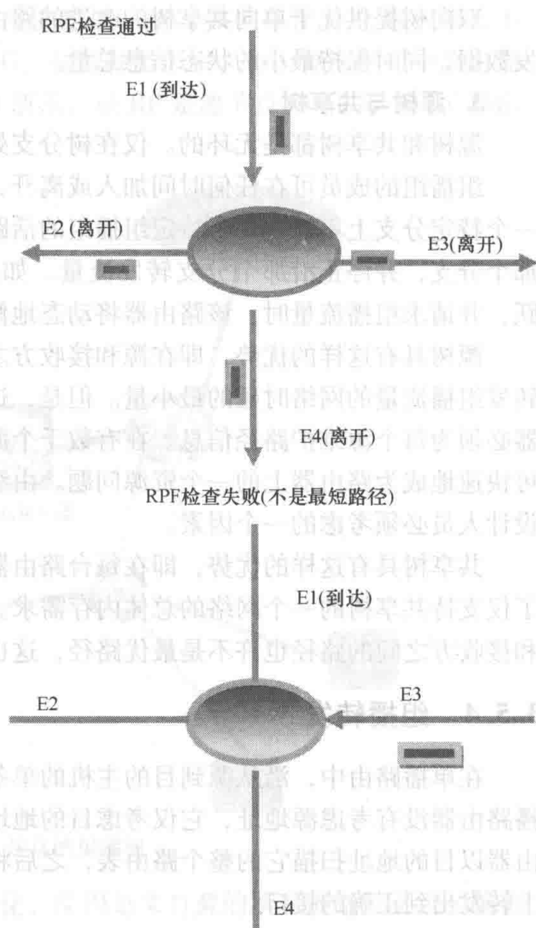


图3.24 RPF检查

RPF检查

当一条组播分组到达一台路由器时，该路由器在分组上实施一项RPF检查。如果RPF检查成功，则转发该分组。否则，丢弃之。

对于沿一棵源树向下流动的流量，RPF检查过程以如下方式工作：

1）路由器在单播路由表中查找源地址，确定该分组是否从回到源的反向路径的接口上到达的，即假定如图3.24所示，到达分组是否从带有IP地址S1的一个组播源到达的。路由器在单播路由表检查是否有一个表项，这里是针对目的地S1的，同样的接口E1将由路由器R使用。如果情况是这样的，那么接口E1是最短路由的组成部分，所以到达分组将被转发到其他组播路由器或在接口E2、E3和E4上的成员主机。

2）如果分组是在回到源的接口上到达的，则 RPF 检查成功，分组被转发。

3）如果在步骤2中的 RPF 检查失效，则丢弃该分组。

2. 反向路径广播

RPB 创建从源到每个目的地的一棵最短路径广播树。它确保每个目的地接收一个且仅接收分组的一个备份。

对于每个网络，选择一台指定的父级路由器。路由器将发送一条 m/c 分组到一个网络，仅当那台路由器是这个网络的指定父级路由器时才这样做。这就降低了到最短生成树的 RPF 中的无环有向图（ADG）。一台路由器可能总是知道邻居中其他哪台路由器有到源的最短反向路径（如果路由算法是基于 DV 路由的）。基于这个事实，选择指定的父级路由器。如果有一台以上的路由器符合条件，则选择具有最小 IP 地址的路由器。

例如，在图3.25所示的网络中，路由器 R1 是网络1的父级 DR，而 R2 和 R3 分别是网络2和网络3的 DR。带有粗线箭头的树表示 RPF，点画线箭头指明 RPB 转发。在这种情形中，路由器 R2 不会将分组转发到网络3。由此，RPB 防止一棵组播分发树中形成环路。

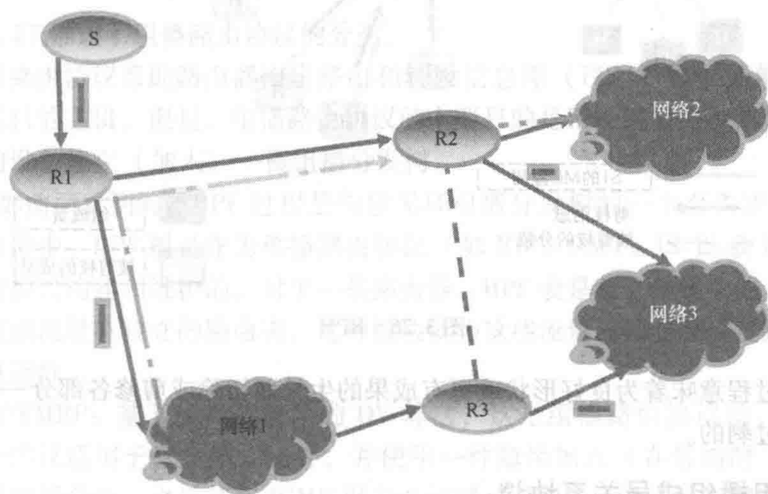


图3.25 RPB

3. 截断的反向路径广播

RPF 和 RPB 广播 m/c 分组。结果是，没有包含 m/c 组的一个网络将接收 m/c 分组，且网络中每台主机的第二层将基于 MAC m/c 地址判定是交付还是丢弃分组。这是没有效率的。

在 TRPB 中，一台指定的父级路由器可判定（通过 IGMP）一个给定组播组的成员是否存在于路由器子网上。如果这个子网是一个叶子网（它没有任何其他路

由器连接到该网络），则路由器将截断生成树。

4. 反向路径组播

当连接到一个网络的一台路由器发现，对 m/c 分组没有兴趣时，它向上行路由器发送一条裁剪消息，从而它可裁剪相应的接口。结果是，上行路由器停止通过那个接口发送针对这个组的 m/c 分组。当一台路由器从一台下行路由器接收到一条裁剪消息时，它终止于将这条消息发送给上行路由器。

如果叶路由器发现其网络之一再次对 m/c（IGMP）感兴趣时，它将发送一条嫁接消息，这将强制上行路由器恢复发送 m/c 分组（见图 3.26）。

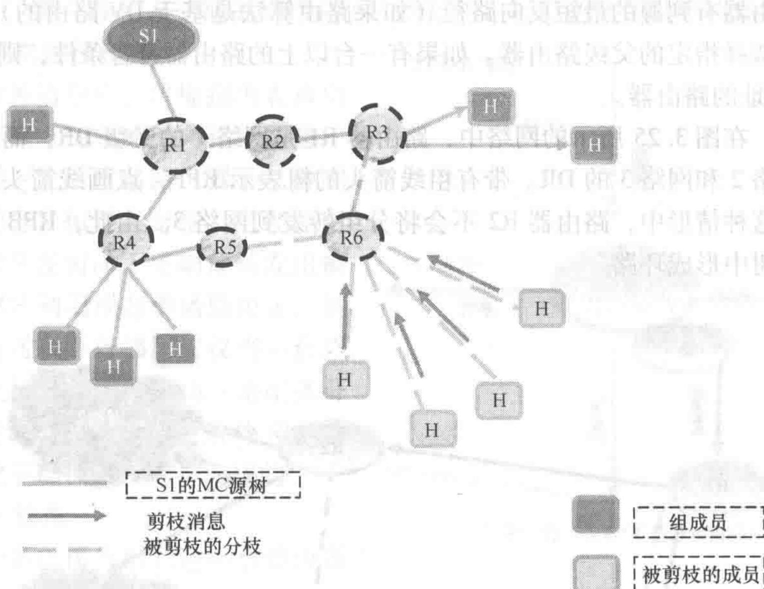


图 3.26　RPM

剪枝过程意味着为良好形状或更有成果的生长而切除或剪修各部分——切除不希望的或过剩的。

3.5.6　组播组成员关系协议

我们看到在组播路由网络中，使用源特定的树或组共享的组播分发树。指存在用于 IP 主机及其直接邻居组播代理之间的组成员关系协议，以便支持组播组的创建、一个组各成员的添加和删除，以及组成员关系的周期性确认。

这些协议为主机和中间路由器动态地加入或离开组播分发树提供一种机制。所使用的主协议被称作 IGMP。使用的当前版本是 IGMPv2 和 IGMPv3（RFC 1112）。它也提供一种报告机制，通知中间路由器和 RP 路由器，组播分发树的早期成员是否仍然是活跃的。如果在一个子网上存在多台组播路由器，那么这些路由器之一是一个 DR，它产生 IGMP 查询和 IGMP 报告。一旦一台组播路由器接收到一条分组，

它将检查组播分发树的至少一个成员是否仍然活跃在与这台路由器关联的树的分段上。如果不再有任何成员，那么它将丢弃分组，之后向上行组播路由器发送一条离开消息。

IGMP 类似于单播网络中的 ICMP。

正常情况下，一台主机将发送一条加入组消息，但可能发送或可能不发送一条离开组消息。但组播路由器仍然向所有主机的组地址 224.0.0.1 发送周期性查询，验证是否仍然存在与组播会话关联的任何组成员。如果没有接收到响应，那么路由器假定该子网不再有组播会话的任何客户。如果接收到一条报告，那么组播分组仍然被转发到该子网。

存在 3 个版本的 IGMP。在原始版本 1（RFC 1112）中，存在一条显式的加入命令，但没有显式的离开消息。相反要使用一次超时（检测节点离开）。后来的版本 2（IGMPv2）提供了显式的加入和离开消息（RFC 2236），支持在一个子网上指定查询的选取，并在一个特定组 ID 上的报告，而 IGMPv3（RFC 3376）主要面向单源组播（SSM）支持优化做出调整。早期版本支持 SSM 和多源组播（MSM）。

3.5.7 组播路由协议

图 3.27 描述了组播路由协议的分类。

单播路由协议帮助路由器构建路由和转发信息库（FIB），以便能够将分组从源路由到目的主机。但是，组播路由协议的主要目的是帮助组播路由器为直接和间接连接的设备建立（加入）一棵组播分发树。

1）如前面看到的，RPF 过程是构建无环组播分发树的一个至关重要的部分。在一些情形中，RPF 树是作为单播路由协议（如 RIP、OSPF、IS-IS 或 BGP）的组成部分而加以构建和维护的。对于一些路由器，RPF 表是独立地维护的。针对单播流量和组播流量有独立的路由表，这种做法就为这些流量而独立地设置路由策略方面支持灵活性。

2）DVMRP：基于从 RIP 衍生的 DV 算法，这是组播路由协议的一个早期版本。这个协议适用于密集模式拓扑，并使用一种隐性加入（在最初时，所有主机都被洪泛组播分组，之后基于 IGMP 报告对组播树剪枝）。它使用一棵基于源的分发树（S，G）。与在 RIP 的情形中一样，对大型网络而言，DVMRP 是不可扩展的。

3）组播 OSPF（MOSPF）：这是 OSPF 的组播扩展。它使用一个显式加入来构建组播分发树，使用密集模式和基于源的分发（S，G）。

4）协议无关组播（PIM）：人们开发了各种组播协议，如 DVMRP、MOSPF 和 CBT。这些协议共有的特征是，它们基于自己的发现机制，构造一个组播路由表。RPF 检查不使用已经存在于单播路由表中的信息。

PIM 使用单播路由表来发现组播分组是否在正确的接口上到达的。RPF 检查是独立的，因为它不依赖于一个特定的协议，它将其判断建立在单播路由表的内容

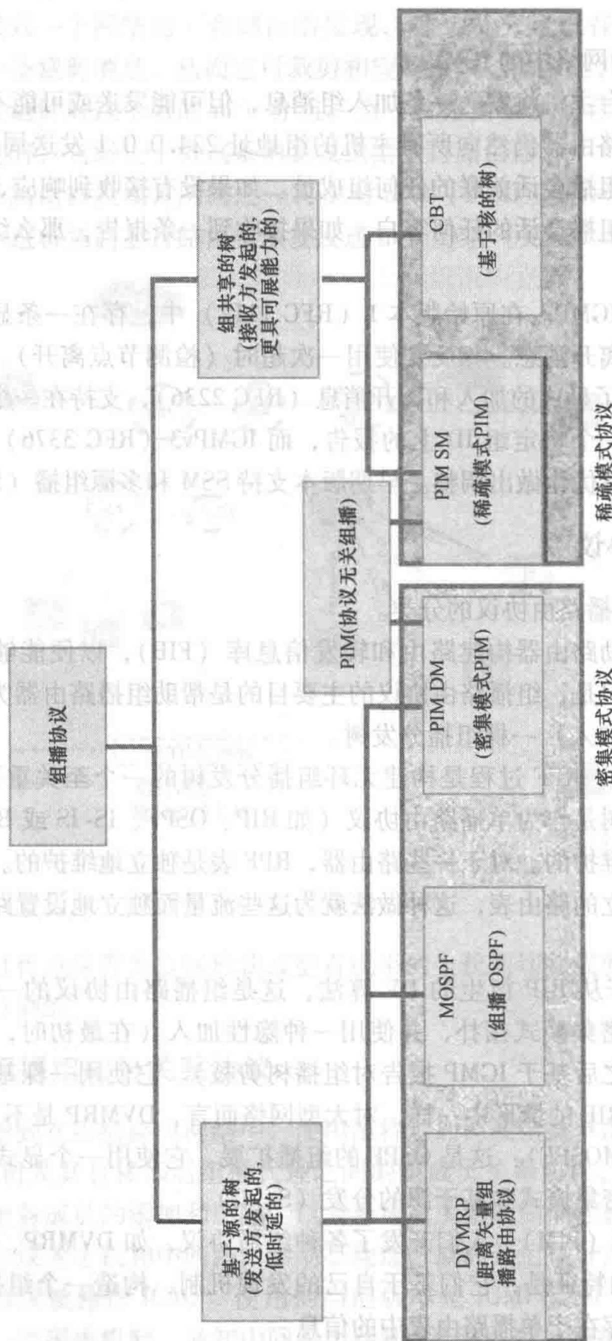

组播协议
基于源的树
(发送方发起的,
低时延的)
组共享的树
(接收方发起的,
更具可展能力的)
PIM(协议无关组播)
DVMRP
(距离矢量组
播路由协议)
MOSPF
(组播 OSPF)
PIM DM
(密集模式PIM)
PIM SM
(稀疏模式PIM)
CBT
(基于核的树)
密集模式协议
稀疏模式协议

图 3.27 组播协议

上。存在几种PIM模式：密集模式（PIM DM）、稀疏模式（PIM SM）、双向PIM（PIM Bi-Dir）和称作SSM的一项新近添加的模式。

1）PIM DM：PIM DM的部署正在逐渐消失，原因是相比PIM SM，它是低效的。PIM DM基于这样的假设，即对于网络中的每个子网，至少对每个（S，G）组播流存在一个接收方。因此，所有组播分组被推送或洪泛到网络的每个部分。不希望接收组播流量的各路由器，因为它们没有那个（S，G）的接收方，沿树向上发回一条剪枝消息。没有接收方的各分支被剪除，结果是一棵源分发树，其各分支有接收方。周期性地，剪枝消息超时，组播流量开始再次洪泛通过网络，直到接收到另一条剪枝消息才会停止。

2）PIM SM：这类似于PIM DM，但它支持主机和路由器的一种显性加入模式，所以各路由器可确定感兴趣的主机，之后构建从各接收方到RP路由器的组播分发树。所以，PIM SM分发树具有形式（*，G）。这被用于因特网协议电视（IPTV）系统，在VLAN、子网或LAN之间路由组播（数据）流。

3）PIM Bi-Dir：这个协议创建一棵双向的转发树。双向组的所有组播路由表项都在一棵（*，G）共享树上。因为流量是可在两个方向传输的，所以状态信息总量可保持最小。路由最优性得以改进，因为流量不会必须在不必要的情况下发往RP。对双向组播组，从来就不构造源树。

4）SSM：在这个协议中，假定在发起一条加入之前，知道一个特定组的源的IP地址。SSM总是在接收方和源之间构建一棵源树。通过一种带外机制了解到源。因为源是已知的，针对源树，可发出一条显式的（S，G）加入，这消除了对共享树和RP的需要。因为不要求RP，就确保了最优路由，流量通过源和接收的最直接路径。SSM是组播网络中的一项最近创新，用于多数新的部署，特别在ISP网络中更是如此。

5）CBT：在特征方面，这类似于PIM SM，但相比PIM SM，是更加高效的。

一些单播协议，如IS-IS（M-ISIS）和BGP（MBGP），也有组播扩展，能够在针对单播分组和组播分组的路由表之间做出区分。

表3.17所示是这些协议的一个快速比较。

表3.17　组播协议的比较

协　议	模　式	加　入	分发树
DVMRP	密集	隐性的	(S, G)
MOSPF	密集	显性的	(S, G)
PIM DM	密集	隐性的	(S, G)
PIM SM	稀疏	显性的，最初之后为（*，G）	(S, G)

对OSPF的组播扩展

如我们所知，OSPF使用Dijkstra的SPF从LSDB推导得到路由信息，LSDB是

从在相同区域中运行OSPF的路由器所发送LSA分组构造的。OSPF将整个AS分成多个区域，具有一个骨干区域（或区域0），所有其他区域连接到该区域。MOSPF是构建在OSPFv2（RFC 1583）之上的。它使用从IGMP报告中推导得到的组成员关系信息，并将这个信息与OSPF LSDB组合来推导组播分发树。与采用OSPF的情形一样，它支持层次结构的和基于区域的路由。因特网被分成各种AS，它们可能被进一步分成各个区域，这就像在OSPF中定义的一样，这也是由组播版本使用的一个概念。

虽然MOSPF算法工作在单个区域，在带有多个区域的一个OSPF网络情形中，各ABR被用作区域间组播转发器，并被用来转发组成员关系信息（作为路由汇总的组成部分），这就使组播分组可跨区域进行转发。在接收区域中骨干路由器的LSDB是由来自各ABR的组成员关系信息更新的。由此，各ABR也包括区域间组播转发器的功能。MOSPF算法添加另一条LSA（称作组成员关系LSA）到OSPF协议中存在的现有LSA之中。这条LSA提供有关节点（订阅到特定组播组）位置的信息，并在LSDB中更新这个信息。

在OSPF的情形中，汇总其他区域的拓扑，并转发到骨干区域，而在MOSPF的情形中，情况未知。因为各LSA总是在区域内洪泛的，不会洪泛到其他区域，所以组成员关系信息也不被转发到其他区域。出于这个目的，ABR作为一个区域间组播转发器。这些路由转发组成员关系信息，由此支持构建一棵组播分发树。之后MOSPF计算SPF树，以源节点为根节点，使用存在于LSDB中的信息利用Dijkstra的算法实施计算。没有导向订阅组播组的一个节点的各条链路被剪枝。如我们所知，在OSPF（和MOSPF）中，每个节点知道任何给定时间整个网络拓扑的状态。只要（S，G）对在相同区域内保持相同，则在相同区域中的所有节点将计算得到相同的组播分发树。MOSPF应需地创建分发树。一旦它从源接收到第一条分组，则它开始构建一棵分发树。给定交付树中的信息，各路由器知道哪些是组播分发的到达接口和外发接口。

通过路由器LSA指明的方法，不管目的地或组为何，MOSPF路由器可接收所有组播分组。是通过打开W（通配符）比特，做到这一点的。这允许MOSPF路由器保持在所有组播分发树上，即使当这些路由器被剪枝时也是这样的（见图3.28）。

在组播路由涉及驻留在不同AS中的一个源和一个目的地的那些情形中，例如在一个MOSPF域和一个DVMRP域之间路由时，它们被看作从DVMRP到MOSPF域的一次路由分发。事实上，与在RIP和OSPF运行在相同域中的情形一样，DVMRP和MOSPF被看作不同域。类似于前面的区域间组播路由情形，AS间组播路由使用一个ASBR作为一个组播转发器和通配符组播接收器。

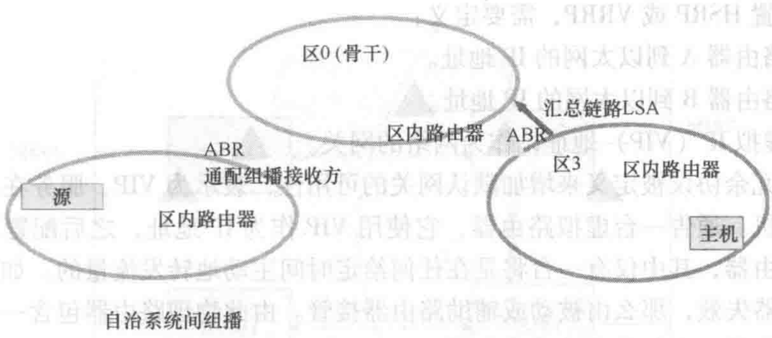


图 3.28　区域间最短路径树

3.6　虚拟路由器和负载均衡

在真实网络中，有必要提供为一台设备接管另一台设备（或在这种情形中的一台路由器）的能力，因为各路由器采用在物理层的各种互联技术工作，且这些技术中的每项技术均可采用特定技术的故障倒换模式加以配置。对于以太网，多数情况下采用 VRRP，这是一种非专有协议，或采用 HSRP，这是一个思科特定的协议。采用 VRRP，也可能针对来自多个厂商的设备进行配置。

VRRP/HSRP（Ayikudy Srikanth，2002）的做法是，配置一台或多台路由器作为一个组的组成部分。一个典型的基本配置如图 3.29 所示。路由器 A 和 B 被配置为处于一个组中，在每个网络上有一个接口。这些路由器之一被设置为主，第二台路由器被设置为辅，当主路由器失效时，由辅助路由器接管流量。基本上而言，这是路由器的一个冗余对，作为一台虚拟缺省网关。在 HSRP 的情形中，主路由器和辅助路由器分别被称作主路由器和被动路由器。

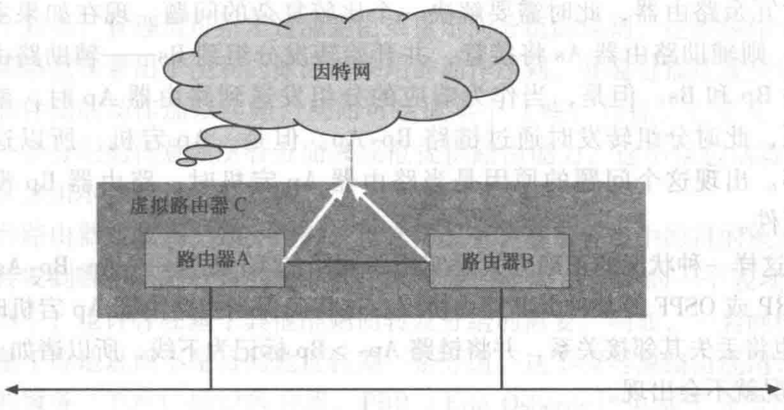


图 3.29　虚拟路由器拓扑的一个视图

为配置 HSRP 或 VRRP，需要定义：

1）路由器 A 到以太网的 IP 地址。

2）路由器 B 到以太网的 IP 地址。

3）虚拟 IP（VIP）地址，作为网络的网关。

这些冗余协议被定义来增加默认网关的可用性，表示为 VIP，服务在相同子网上的各主机。通告一台虚拟路由器，它使用 VIP 作为 IP 地址，之后配置两台或多台物理路由器，其中仅有一台将是在任何给定时间主动地转发流量的。如果当前活跃的路由器失效，那么由被动或辅助路由器接管。由此物理路由器包含一个主或主动路由器和另一台备份或被动路由器。

这些路由器可被指派优先级，以便首选地选取一台路由器作为主或主动路由器。VIP 活跃在其优先级较高的无论哪台路由器上。优先级默认地被设置为 100，但该值的范围是从 0 到 255。

所有 HSRP（Jeff Doyle）分组被设置为 1 的存活时间（TTL），所以这些分组从来不会离开本地网络分段。HSRP 分组在 UDP 端口 1985 上被发送到组播地址 224.0.0.2。

当运行 HSRP 的一台路由器初始化时，它发出 HSRP hello 分组，以便确定在相同网络分段上是否有运行 HSRP 的路由器。如果找到一台以上的路由器，则各路由器协商以确定主动的或主路由器。如果由于相等的优先级值而导致平局，那么具有较高 IP 地址的路由器成为主动路由器。

在 HSRP 中，当主路由器重新上线时，有必要强制辅助路由器（它已经作为主路由器）放弃作为主路由器的地位。如果有两台以上的路由器参加，则一旦主动和待命路由器的选取完成，那么其他路由器既不是主动的也不是待命的，直到待命路由器成为主动路由器时才可能发生变化。

当在两个网络 A 和 B［跨广域网（WAN）连接的，见图 3.30］的边缘处定义两对冗余路由器，此时需要解决一个比较复杂的问题。现在如果主路由器 Ap 宕机，则辅助路由器 As 将接管，并开始转发分组到 Bs——辅助路由器虚拟路由器对 Bp 和 Bs。但是，当作为响应的分组发送到路由器 Ap 时，就出现了主要问题，此时分组转发时通过链路 Bp-Ap。但是，Ap 宕机，所以这些分组将被丢弃。出现这个问题的原因是当路由器 Ap 宕机时，路由器 Bp 没有注意到这个事件。

处理这样一种状况的正确方式是实施跨网络的 L3 配置——Ap-Bp-As-Bs，使用像 EIGRP 或 OSPF 等某种 IGP 路由协议。这将确保当主路由器 Ap 宕机时，主路由器 Bp 也将丢失其邻接关系，并将链路 Ap->Bp 标记为下线，所以诸如上面一种情况的状况就不会出现。

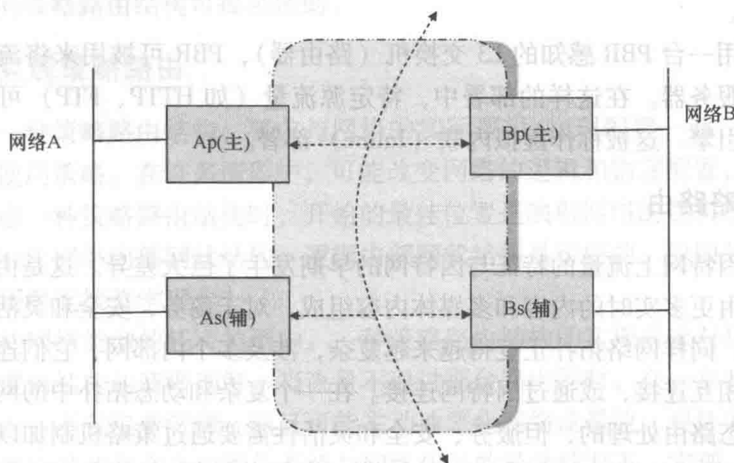


图 3.30　跨网络的虚拟路由器配置

3.7　基于策略的路由

3.7.1　引言

基于策略的路由（PBR）（Net）是这样一项技术，基于由网络管理员设置的策略做出路由决策。

正常情况下，所有路由都是目的地驱动的，即当路由器实施分组处理时，它仅关注于分组的目的地址，它被用来就在哪里和如何转发分组或丢弃分组做出一项决策。但是，在一些网络设置的情形中，也许有必要以不同方式路由分组，这取决于目的地址甚至分组首部中的其他字段，如 IP、源地址、TCP 或 UDP 端口或净荷。为达到这个效果，管理员可指定过滤路由或重定向分组的规则。这被称作 PBR。

一个策略意味着出于便利的原因而采用的动作序列，并通过描述或预先描述一个规则集和伴随的动作加以实施，从而可达成某个（些）目标。它定义一个规则集，基于一条分组的任意或所有方面来规范提供路由能力。这不仅包括首部信息，而且包括在分组本身内包含的数据。

当一台路由器接收到一条分组时，正常情况下，基于分组中的目的地址，它决定要将之转发到哪里，之后目的地址被用来查找一个路由表中的一个表项。但是，在一些情形中，也许存在基于其他准则而转发分组的需要。例如，一名网络管理员也许希望基于源地址而不是目的地址转发一条分组。这不应与源路由混淆，其中以不同类型的服务（TOS）标记各分组。PBR（Eric Osborne）也基于分组尺寸、净荷的协议或一个分组首部或净荷中存在的其他信息。这允许将来自不同源的分组路由到不同网络，即使当目的地相同时也可这样做，当互联几个私有网络时，这种做

法是有用的。

通过使用一台PBR感知的L3交换机（路由器），PBR可被用来将流量重定向到一台代理服务器。在这样的部署中，特定源流量（如HTTP、FTP）可被重定向到一个缓存引擎。这被称作虚拟内联（Inline）部署。

3.7.2 策略路由

如今，因特网上流量的特征与因特网的早期发生了巨大差异。这是由商务需要驱动的，并由更多实时的内容和多媒体内容组成。对于商务，安全和灵活性成为主要担忧问题。同样网络拓扑正变得越来越复杂，涉及多个内部网，它们连接到因特网，并内部相互连接，或通过因特网连接。在一个复杂和动态拓扑中的网络路由是通过使用动态路由处理的，但服务、安全和灵活性需要通过策略机制加以处理。同样由于因特网上流量的变化特征，有选择地区分和路由流量就变得必要了。出于这个目的，基于QoS的路由也成为重要的。

各种因素都收敛到PBR方案，这正如在如今的网络中看到的情况。因特网流量的变化特征，导致数据流分优先级，导致基于QoS路由的开发。一些其他考虑，包括对安全的需要、流量分离、不同组织的不同商务需求等。

但是，策略路由结构的主要驱动力可被看作是QoS。传统IPv4流量转发是基于尽力而为模型的。一个ISP可确保一名消费者得到某种得到保障的网络服务水平，其中消费者为之支付了一项优惠的（Premium）服务水平协议（SLA）。在IPv4网络下，这种机制取决于网络分组的路由和排队，取决于与所提供QoS服务相关联的TOS标记。多数QoS机制基于区分服务（DiffServ），它指定IP数据报中的TOS字段，定义排队律的各种等级。

分组中的TOS标记可被用来提供分组交付的某些保障。单独使用简单的排队，则仅有的保障是被标记的排队将得到总可用带宽的一个确定百分比。采用基于TOS标记的策略路由，在拥塞避免和优先分组路由的方法中可添加新方法。排队结构，通过提供对现有有限带宽连接的中介（Mediated）优先访问，支持对分组数据流的更好服务。

3.7.3 策略路由结构

策略路由也被称作智能路由，这是由于作为路由的组成部分，中间设备所要求的智能才这么称呼的。参与网络基础设施的任何设备，也可以是智能或PBR的组成部分。更常见的情况是，PBR是在核心网络中实现的，作为核心网络所实施路由的组成部分。

随着诸如流化音频和视频等服务被引入企业网络，对网络内分配资源和经济地使用资源，出现了巨大的需求。通过重新设计和优化网络流量，使用策略路由，返回一个高得多的成本节省，其中实现了巨大的效率。在许多前沿方面，一种仔细设

计和实现的策略路由结构可提供辅助。

3.7.4　实现策略路由

实现一种策略路由结构，要求与网络的实际逻辑/物理配置一起，考虑所有现存的网络使用策略。在许多情形中，可能改变网络的逻辑和物理配置，以有利于实现。当考虑一种策略路由结构时，开始的最佳位置是映射网络的逻辑结构。这个逻辑地图将给出网络内部网状结构。逻辑内部网状结构是重要的，原因是如今多数网络仍然采用单连接哲学理念。

在这些网络类型的任一类型中，一种策略路由结构的实现要求对目标的仔细分析和实际逻辑结构的清晰理解。当流量不通过那台路由器时，在一台叶路由器上实现策略路由，不仅浪费资源，而且可能主动地恶化网络流量流。更坏的情况是，在不理解分组穿越路径和桌面操作系统与网络相关的异常情况下，实现一种策略路由结构，可能使网络崩溃。

策略路由结构被用来实现安全性、网络和路由策略。当考虑一种策略路由结构的实现时，必须理解整个网络和网络操作的范围。理解网络的用途和穿越网络的各协议的操作，对于设计一种良好的策略路由结构是至关重要的。

在许多商用路由器中，使用 route maps 实现 PBR，而在 Linux 环境中，通过使用诸如 ipchains 和 iptables 等命令加以实现。

PBR 的一些优势有：

1）像 ISP 的各组织机构路由如下流量，这些流量源自不同的源地址，源自不同的用户组（应用）。

2）各组织机构可基于边缘网络中 IP 分组首部的 TOS 字段的值，区分流量。采用这种方式，通过核心网络的流量数据流可被汇聚，并以相当快的速度移动。

3）一个组织机构也可将与一个特定活动相关联的块式流量，指导在短时间内使用一个较高带宽、高成本链路，并保持基本的连通能力。

4）可定义策略，基于各种流量特征将流量分配在多条路径间。PBR 使路由器能够在进入接口处评估所有流量，这里使用为该接口配置的路由映射。策略规则被定义为 permit（允许）或 deny（拒绝）语句的一个组合体。一条 deny 语句意味着匹配准则的分组通过正常信道发送，而一条 permit 语句意味着，如果分组匹配所有准则，那么应用所有相关的命令。如果分组不匹配路由映射中的任何规则，那么分组通过正常路径被转发。例如，为丢弃匹配某个准则的各分组，作为最后的规则，各分组被路由到接口 null 0。

5）与 PBR 相关的第二个控制集指各种过滤器。各过滤器可被用来在对端接口之间或路由协议之间路由信息，如在分配路由时的情况。这些也可被用来停止发送路由更新到邻居，过滤路由以便防止出现路由环路等。

过程如图 3.31 所示。

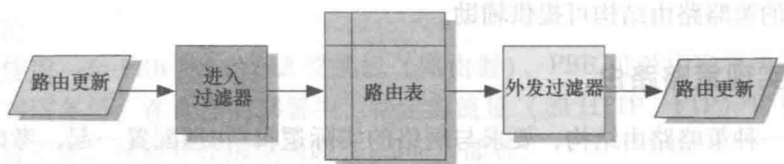


图 3.31 PBR 过程

由此 PBR 用于：

1）过滤在路由更新中发出的路由。

2）过滤在路由更新中接收到的路由。

3）对一个路由度量指标施用一个偏移。

4）评价一个路由信息源的可信性。

重新分发路由：在一个 AS 间网络中，从一个 AS 到另一个 AS 的路由及相反方向的路由，由于路由前缀、路由度量指标和可能的不同路由协议中的差异，需要实施“变换”。这个过程被称作路由分发。当在两个方向发生路由分发时，被称作双向分发（见图 3.32）。

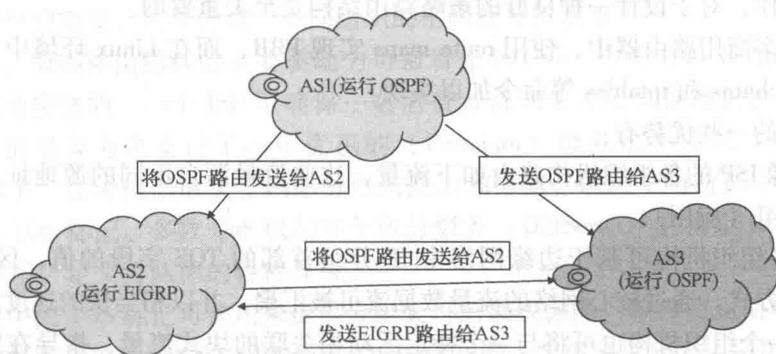


图 3.32 PBR

PBR 在域间环境中的 BGP 路由、路由分发和客户就绪准备方面扮演一个极端重要的角色。PBR 机制是主要作为能够控制从一个 AS 到另一个 AS 分发路由所需的结果才出现的。在一个域间环境中，从一个邻接 AS 接收到的路由公告可能有这样的 IP 前缀，即接收 AS 可能不希望处理或以不同于发送 AS 的方式进行处理。同样各种商务协议也影响路由被赋予优先级。在三个阶段中发生 PBR：

1）确定策略并将这些策略安装在路由器中。

2）当发生路由更新时，应用策略来更新 RIB 和 FIB。

3）当实际数据分组到达时，它匹配在一条或多条策略中规范的规则时，应用通过 FIB 规范的动作。

也已经定义了一种通用的路由策略规范语言（RPSL，RFC 2622）来规范策略，

这些策略可在异构厂商的设备之间进行交换。但是，多总比没有好，针对这个目的，客户似乎喜欢使用厂商提供的平台特定工具。

对于多个 AS、多个边界路由器和多种策略，整体的系统性行为可变得非常复杂，并可能导致振荡或不希望的行为，可导致网络的一些部分变得不可达。

3.8 路由器和交换机：平台架构

一台路由器的基本功能可被分类，如图 3.33 所示。

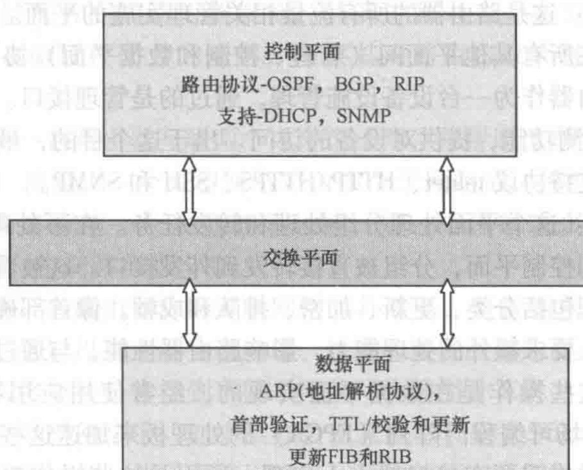


图 3.33　一台路由器的功能组件

1. 处理

这包括路由路径计算、路由表更新和分组的过滤，取决于策略设置。

2. 分组转发

分组转发的功能包括：

1）首部确认：检查每条分组的版本、协议字段、首部长度和首部校验和，以便检查分组的有效性。

2）路由查找：每条进入 IP 分组的目的 IP 地址实施交叉检查，以便确定分组的目的地是当前路由器还是另一台路由器。如果分组指明是本地路由器，则由处理器加以处理。如果其目的地址是另一台路由器，则以目的地址检查路由表项，且分组通过合适的接口加以转发。

3）分组更新与分组校验和更新：也处理各种分组字段更新，其中包括 TTL 值。如果分组要被丢弃，那么一条 ICMP 不可达消息被发送到源主机或路由器。每次更新任何分组字段时，也需要更新分组校验和。

4）分段：取决于 MTU 尺寸，在各种链路上对分组实施分段和重组，可能是必

要的。

3. 辅助功能

这些功能包括分组的基于 QoS 的调度和流量优先级处理，基于策略的分组过滤、网络管理操作和统计数据的更新。

封装上述路由器功能的逻辑组件可被分解为三种组件：

1）控制平面：控制平面处理一台路由器的所有主要功能，包括所有路由协议的路由计算、分组检查、处理和分类。控制平面的一部分也处理管理功能，如统计和简单网络管理。要求任何处理的所有分组都是通过控制平面的路径（慢速路径）发送的。

2）管理平面：这是路由器的所有流量相关管理实施的平面。它提供一个网络所需的功能，并在所有其他平面间（管理、控制和数据平面）协同功能（执行）。它也被用来将路由器作为一台设备设施管理，通过的是管理接口。这层通过命令行界面（CLI）和监测功能，提供对设备的访问。出于这个目的，最常见的是，为配置和监测设备，支持协议 telnet、HTTP/HTTPS、SSH 和 SNMP。

3）数据平面：这个平面处理分组处理和转发任务。在不处理分组的情况下，通过将分组路由到控制平面，分组被直接转发到外发接口。这被看作是快速路径路由。分组转发处理包括分类、更新、加密、排队和成帧。像首部确认和目的地路径决策判定等操作，要求额外的处理能力，影响路由器性能。与通过控制平面路由分组的做法不同，这些操作是在数据平面实现的，经常使用应用特定的集成电路（ASIC）或基于现场可编程门阵列（FPGA）的处理板来加速这些操作，并避免使 CPU 过载。一些操作需要在接口线卡上实现，原因是这些操作要求网络访问，如 ARP 处理。

路由器组件和架构：一个物理观点

一台路由器由图 3.9 所示的基本组件组成：

1）多个网络接口卡（或线卡）与附接的网络接口。

2）处理模块或监控卡。

3）内部交换结构。

最常见的是，这些组件是使用实现比较快速处理的一个或多个 ASIC 构造的。在到达接口处接收分组之后由处理模块处理，再通过交换结构转发到外发接口。诸如接口配置和统计信息收集等其他功能是在管理或控制平面中处理的。

逻辑上而言，路由器架构或功能可被分成三个组件，经常称作平面。这些平面如下：

1）控制平面：这个平面负责与其他路由器一起，参与执行路由协议、接收和广播（或单播或组播，取决于协议）更新的网络拓扑信息，并构建或更新路由表。同样各种其他功能，如丢弃分组或分组的 QoS 服务，也是在控制平面中实施的。

2）数据或转发平面：这个平面负责实际转发流量。来自路由表的信息被用于

构建转发表，该表被用来实际上将分组转发到外发目的地。在图 3.9 中，正常情况下，监控卡实现控制平面，而接口或线卡实现数据平面。多数情况下，线卡使用三路（ternary）内容可寻址内存（TCAM），进行转发路径的快速存取。

如我们所看到的，就在网络中转发分组和管理流量而言，路由器在网络中扮演一个关键角色。本节将简短地看看这台设备的内部机理（Innards）。随着通信基础设施的速度增加和各种计算设备（包括移动设备）处理能力的增加，对通过因特网（由此通过路由器）传输的流量总量和速度正导致路由器架构也发生演进，并在分组处理和操作方面变得更加快速、更加功能强大和高效。本节将简短地讨论硬件和软件层次方面的路由器架构、其演进和未来。本节将讨论路由器所需的功能，得到基础硬件平台的理解，考察其对路由协议实现的影响，之后考察未来方向。

基本上而言，一台路由器互联两个或多个子网，并帮助选择性地在这些网络之间传递（双向地）流量。一个大型网络可由多个子网组成，这些子网通过多台路由器连接。一个网络跨越一个大型地理区域，如一个国家或多个国家（如因特网），也需要由多台中间路由器连接。

在这种情形中，各路由器将需要就网络拓扑状态、流量拥塞等方面交换信息。基本上而言，它是针对以高速的分组处理、过滤和转发的目的而进行定制的。一般而言，可利用带有多个接口连接到不同子网的一台通用计算机，并将之变成一台路由器，它可部署在小型网络中，如小型办公室家庭办公室（SOHO）网络。基本上来说，这正是像 Vyatta 的一些公司使用基于开源软件的产品所做的。路由器也不得不处理多种类型的接口，如以太网、光纤、无线、ATM、PSTN 等。因为它可被连接到多个网络，利用这些技术中的多种技术，在物理层和 L2 层进行流量处理。在 L3 层，路由器也必须支持多种路由协议，如 RIP、OSPF 和 BGP。多数情况下，路由也被称作“L3 交换”。这具有更多历史方面的内涵，因为 L2 交换是从诸如集线器的 L2 层设备衍生出来的。后来这些交换机有一些其他 L3 层功能添加于其上。通用术语“交换”指 L2 层分组处理。

如前面看到的，一台路由器的功能可被分成三层或部分：控制平面、转发平面和用户流量平面。

一台路由器的容量，正常情况下，规范为每秒分组数（pps）。取决于路由器架构，这些具有不同的性能曲线，这些曲线取决于分组尺寸和所要求的分组处理量。多数情况下，路由器将为分组处理提供多条路径，取决于采用或指定哪条路径，性能将发生变化。在路由器性能中的一些瓶颈有路由器处理、分组处理（如所要求的分组检查和过滤）和路由器架构。

正常情况下，取决于一台路由器的容量（就它所能处理的流量总量而言），路由器可被分类为核心路由器、接入路由器或边缘路由器。一台核心路由器被设计为在因特网的骨干（两个或多个层 1 ISP 的 NAP）以一个非常快速的速率处理流量。

它必须能够支持多种类型的电信/通信接口和多种路由算法：PBR，使用访问控制列表（ACL），基于各种参数过滤路由和流量的能力。一台边缘路由器驻留在骨干网络的边缘，并连接到核心路由器。由 ISP 使用的各路由器倾向于使用 BGP，与其他 AS 中的各路由器交换路由信息。一台用户型边缘路由器或客户边缘（CE）路由器位于用户网络的边缘。

如在前面因特网架构一节中看到的，因特网核心本质上是核心路由器集合，它们在各 NAP 位置处连接各层 1 ISP。也存在提供间（interproviding）边界路由器，交换路由信息。这些路由器正常情况下使用 BGP 路由协议。出于转发一条分组的目的，正常情况下，一台路由器检查源和目的 IP 地址。针对分组流（如多媒体或 VoIP 分组）的基于 QoS 调度，一台路由器也查看一下 L2 首部信息。一台路由器不针对通过该路由器转发的任何分组维护任何状态信息。当到达分组速度快于路由器能够转发分组的速度时，它管理流量拥塞。路由器也基于各种策略实施分组转发，称作 PBR。在多数现代路由器中，存在多个 ASIC 引擎，构造它们是为了能够以硬件而不是以软件实施许多这样的功能。基于 L2 层信息的分组转发被称作交换。

路由器设计的一个更具决定性的因素是它在其中部署的网络类型：

1）骨干网络：这些是核心路由器。它们被用来连接 ISP 网络，并在 T bit/s 速度下实施转发（所以实施的是光互联）。这些需要是非常可靠的、容错的和热交换的。性能、扩展性、高可靠性和可用性是极端重要的特征。核心路由器可连接到大量接入路由器，通常是数百量级。

2）接入网络：在这个网络层中的路由器合并来自多个客户的流量，并将流量驱动进入核心网络。这些需要支持多个物理接口和多个高速（OC-12 +）连接。

3）企业网络：正常情况下这些由用户端客户组成，包括机构（Institutions）和驻地客户。

1. 中心式路由/共享总线架构

这个架构（见图 3.34）使用一个通用处理器，采用一条共享总线连接到多个接口。共享总线被用来互联所有接口，而分组转发、处理等所有功能是由处理器实施的。随着分组到达进入接口，这些分组通过共享总线被发送到处理器进行处理，在处理之后，这些分组第二次在共享总线上被发送到外发接口。这不是一种可扩展的架构，但可采用现成商用组件实现。这是诸如 Vyatta 等公司所实际遵循的战略，使用开源软件和现成商用硬件构造路由器产品。

虽然这种架构是比较廉价的，且实现起来简单得多，但在这种架构中存在许多限制因素：

1）中央处理器成为一个瓶颈，原因是它不得不处理从进入接口到外发接口通过的每条分组。

2）路由表查找和分组转发是内存密集型的操作。

将分组从进入接口移到外发接口，要求共享总线存取两次，所以限制了路由器

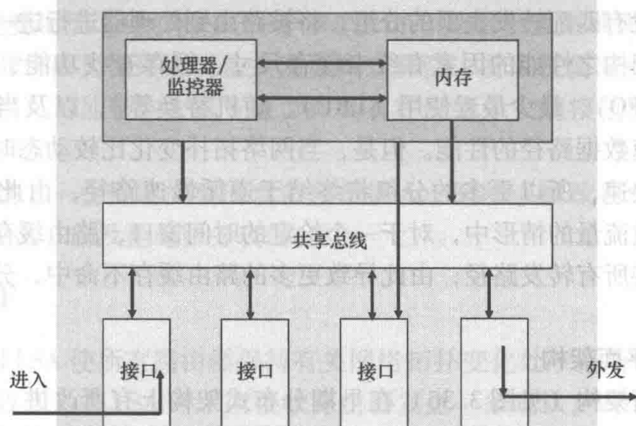


图 3.34　共享总线路由器架构

吞吐量和性能。基本上而言，在这个架构中存在许多限制因素。

2. 分布式路由器架构

这个架构（见图 3.35）克服了前面看到的共享总线架构的一些限制。在这种情形中，分组处理部分地是在接口（或线卡）中完成的。

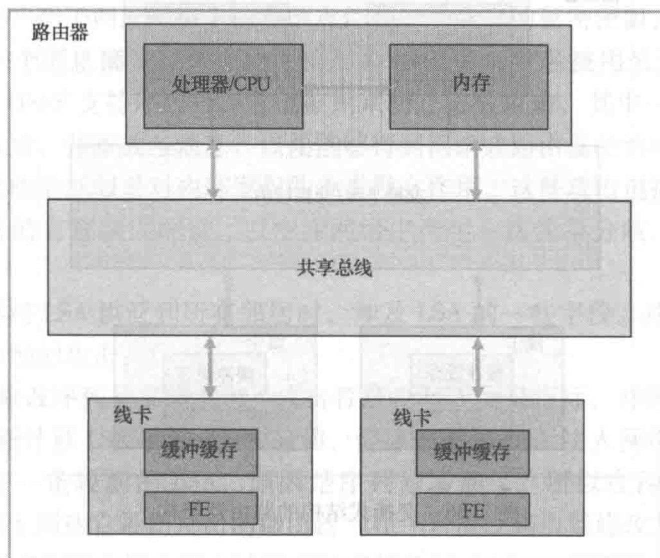


图 3.35　分布式路由器架构

每个线卡有其处理器，并缓存发送和接收缓冲。

这限制了进出主存的拷贝分组（速度）和共享总线被存取的次数。线卡也缓存用于分组转发的最新路由。采用这种方式，目的地为其他路由器的分组，多数通

过进入接口被转发到外发接口，由此降低了处理器、内存和共享总线上的负载。在线卡总是缓存没有匹配转发表项的分组，将被路由到处理器进行进一步处理。

影响这个架构之性能的因素有线卡缓存尺寸、缓存查找功能、缓存维护策略[先入先出（FIFO）、最少最近使用（LRU）、随机替换等]，以及当通过一个共享总线路由时慢速数据路径的性能。但是，当网络拓扑变化比较动态时，路由缓存中的变化要更加快速，所以更多的分组将终结于遵循慢速路径，由此影响路由器性能。同样在繁重流量的情形中，对于一个给定的时间窗口，路由缓存将不太会提供足够的内存保持所有转发路径，由此导致更多的路由缓存不命中，分组所走的慢速路径次数增加。

3. 交换式平面架构

交换式平面架构（见图 3.36）在早期分布式架构上有所改进。这个架构引入一个完整的路由表和转发数据库作为线卡的组成部分。当网络拓扑动态时，就吞吐量和网络抑制能力（Resilience）而言，这有助于进一步降低慢速路径的使用并改进性能。如我们看到的，在高速流量窗口中，采用慢速路径的分组，可能增加并由此影响路由器的性能。在这种情形中，共享总线可成为一个主要瓶颈。通过以交换式互联结构替换共享总线，分组在进入和外发线卡之间传递，或从进入线卡到处理器模块，之后到外发线卡，而没有任何时延（线速）。

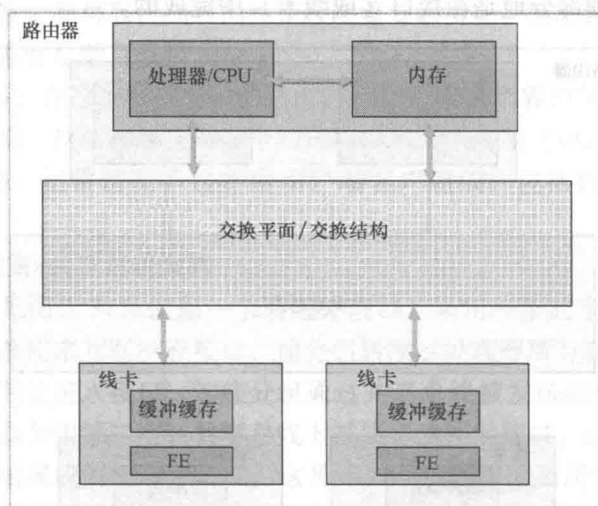


图 3.36　交换式结构的路由器架构

3.9　安全管理

在交换和路由器中经常逃出考虑的领域之一是安全领域。这里将简短地考察一下与主要协议（像 OSPF 和 BGP）相关的安全问题，原因是这些协议是在因特网中

正在用的主导路由协议。

OSPF是主要用于AS网络内的一个IGP，BGP是用于独立域或AS或ISP之间路由的一个EGP。因为网络协议是并发分布式算法，所以任何安全弱点（通过一次攻击可被利用）可能影响大量节点，并具有在网络中大规模范围上中断流量的可能性。

在一个网络上一些特定类型的攻击包括插入、删除或修改在中转中的分组或LSA，中间人攻击，以及拒绝服务攻击。

3.9.1 OSPF

OSPF使用LSA使所有路由器保持有关网络拓扑变化处于最新状态。在从一个邻接路由器接收到一条LSA时，由接收路由器发送一条确认，该LSA也被转发到当前路由器的所有其他邻居。

在OSPF安全中的主要问题之一是内部人员攻击问题，这在协议设计中没有给予正确的考虑。OSPF是一个复杂的协议。所以实现一个完备的安全方案也是困难的。任何安全方案，当没有被合适地实现时，将不可避免地导致协议中的严重安全瑕疵。

在OSPF中，要求认证路由器之间的所有交换。非常常见的情况是，认证涉及以明文传输一个口令，这可通过窃听而加以捕获。在后来的版本中，实现密码学方式的认证。在一个子网或网络上的各节点使用一个共享的秘密密钥，这被用来为每条LSA产生一个消息摘要。这个消息摘要5（MD-5）签名被用来认证一条分组。如我们所知，OSPF支持序列号，它也被用来防止重放攻击，其中一名攻击者捕获来自网络的流量，并重放各消息，以便能够得到网络或路由器的访问。

但是，这些措施似乎对内部发起的攻击没有作用，这些攻击包括一个已知缺陷的利用或设备的有意错误配置，以便在网络中产生一次流量分歧（Diversion）或中断。

当路由器将LSA洪泛到所有邻居时，涉及LSA的一些字段。这个步骤可被用来发起各种类型的攻击：

1）通过修改序列号实施攻击：攻击者修改LSA度量指标，并将LSA序列号加1。同样，重新计算LSA和OSPF校验和，经修改后的LSA注入网络。现在其他路由器假定这是一条较新的LSA，原因是序列号增加了，所以在网络中传播这条LSA。当该LSA到达它最初发出的路由器（在一台流氓路由器修改并重新注入流量之前）时，假定该路由器会清空或修正这条修改过的LSA，原因是它将认识到在被修改的LSA中指定的各项资源不是它所具有的资源。但是，人们观察到，这种防御机制可能无效（Nullified），方法是通过使用幽灵（Phantom）路由器，在网络中可重复地或周期性地注入流氓LSA。同样，存在一些OSPF实现，它们未必清空这些LSA。

2）通过修改 MaxAge 实施攻击：流氓路由器可修改 LSA 年龄字段为 MaxAge 值，之后在 LSA 被修改之前计算当时的校验和。当修改过的 LSA 被注入流量中时，因为 LSA 的年龄被设置为 MaxAge，则所有路由器将从 LSDB 中清除实际有效的 LSA（流氓路由器从有效的 LSA 产生修改的 LSA)。

3）MaxSequence 号攻击：遵循与前面的攻击采用的一种类似方法，其中攻击者将序列号设置为 MaxSequenceNumber，并将该 LSA 注入网络。因为序列号被设置为 MaxSequenceNumber，所以各路由器假定这条 LSA 为“最新的”LSA，并使用它替换较早的有效 LSA。当这条 LSA 到达通告原始 LSA 的原路由器时，假定源发路由器会清除这条 LSA。

3.9.2 BGP

设计 BGP 路由协议，使各 AS 能够交换路由或将之用于域间路由。在设计期间的假定之一是，各 AS 是相互信任的，就 BGP 协议设计中安全而言，不需要做特殊考虑。假定网络是被信任的网络。BGP 不保护消息的完整性或来源，就可达性信息而言，没有验证 Ass 的权威性。

因为 BGP 在 TCP 连接之上交换信息，如果 TCP 连接被截获，则可在 BGP 上发起各种类型的攻击，诸如中间人、重放或拒绝服务攻击。各路由器可被错误配置为产生假的通告，或定时器可被增加以产生更频繁的更新。

一些更常见的攻击有：

1）窃听：通过侦听在链路上传递的数据，攻击者可了解策略和路由，这些可能是敏感信息。

2）重放：消息可被记录，并发回到原接收方，导致服务的中断。

3）消息插入或修改：通过伪造消息，一名攻击者可将不良路由注入路由表中。

4）中间人攻击。

5）拒绝服务攻击。

BGP 的安全版本（S-BGP）提供使用一项公开密钥基础设施（PKI），对路由更新实施签名。地址分配证书认证由一个特定 AS 拥有的地址范围，而管理系统证书验证 BGP 发言者（Speaker）确实被授权将路由分发到其他 AS。接收方可使用标准 PKI 基础设施进一步验证这个消息。

3.10 电信和公众网络：交换和路由

许多电信和公众网络是由有许可证的服务提供商运营的，它们的核心商务是提供电信服务。电信网络采用一种层次结构拓扑。提供骨干路由，连接每个交换中心，从最低层到最高层，通过各种中间层次的交换中心实施连接。

在电信网络中，编址规划、路由和计费是紧密相关的活动。对所有三项活动而言，预期电话号码的编址规划是一致的。编址、路由、缴费、传输和信令的国家规划，形成互相关的一个标准集合，这些标准监管国家网络和本地网络的规划以及与国际网络的相互作用规划。

电信网络的编址规划在本质上是层次结构的，以便可处理各种呼叫方案，如直接远距离拨号（DDD）、用户中继拨号（Subscriber Trunk Dialing，STD）和国际用户拨号（ISD）。

电话网络一直是在用的主要 PSTN。非常常见的是，所提供的基本电话服务被称作老式电话服务（POTS），以便能够将之与由服务提供商所提供的有更高附加值的服务区分开来。ISDN 过去被看作是对 PSTN 的另一项演进性的步骤，它为话音和数据服务提供信道。公众服务网络有其他各种变形。例如：

1）移动电话网：这些是蜂窝或移动电话网，多数是区域性的或国家范围的，它们连接到 PSTN，为的是实现长距离和国际连接。

2）寻呼网络：在移动网络的演进过程中，寻呼网络扮演了一个重要角色，但现在则扮演一个不太重要的角色。这些是用于单向消息的消息网络。在最近时间以来，诸如电子邮件和传真的基于因特网的网络服务，多数情况下替换了这些服务。

3）公众数据网：在因特网大范围扩散之前，这些网络为客户提供点到点、电路交换或分组交换网络连接。非常常见的是，这些网络使用 X.25 协议实现连接，是基于 ITU-T 的 X 系列协议建议标准开发的。针对数据通信目的，这些多数情况下是由企业客户使用的。依据使用情况，对客户收费。在最近时间以来，诸如电子邮件和传真的基于因特网的网络服务，在多数情况下替换了这些服务。如今移动网络在移动网络上为数据流量提供通用分组无线服务（GPRS），而在有限距离内为无线数据服务提供诸如 Wi-Fi（802.11）和 WiMAX（802.16）的无线服务。

智能网（IN）为数字电话网提供了附加的智能能力，像灵活的呼叫路由、重定向和通知。在这样一个网络中，用户电话号码可以是物理的或逻辑的，并提供通知一组特定服务（诸如 800 或 900 服务）的一种方式。取决于所拨的号码和其他参数，如由用户完成的时间和数字选择，实际呼叫可被路由到各端站，它们为用户提供一种特定服务（或多项服务），例如将紧急号码路由到位于一个城市不同部分的各运营商。

诸如呼叫转移、呼叫等待、自动回呼、缩略拨号和呼叫屏蔽（呼叫者 ID）等服务，帮助服务提供商增加收入和网络利用率。现代电话交换机，与早期电子机械式交换机不同，具有巨大容量，能够支持进行中的数千名用户和同时呼叫。

信令机制支持客户端设备或交换机建立、维护和终结网络中的呼叫。信令由特定消息组成，这使端点可沟通它们的状态，并依据从另一端接收到的信息改变状态，如摘机或挂机状态和拨一个号码；多数呼叫建立涉及一系列交换机。依据连接被使用的时长，服务公司可对用户收费。电信网络中的信令是一个复杂过程，涉及

多台交换机，甚至多家运营商，诸如一个移动呼叫（从一个国家发起到另一个国家中的一个移动号码）。用户A必须通过本地移动电话网络连接到一个PSTN，这将允许呼叫被转发到目的国家中一个PSTN的一台交换机，这将通过目的PSTN到目的移动电话的移动网络发起一次呼叫。在这种情形中，在建立呼叫、对呼叫时长缴费和划分收入以及维护呼叫中，涉及多家运营商。

诸如上面的一项能力（从一个电话号码到另一个国际电话号码的呼叫）是可能的，这是由于为世界上每名电话用户提供的唯一标识做到的。电话服务提供商使用一种层次结构编址方案，其中国家代码在最高层。依据E.164方案，完成国际电话编号。

一个国际接入号码或前缀被用来通知网络它是一个国际呼叫。国家代码，从1个数字到4个数字，识别目的国家，如91是印度，65是新加坡等。

其他电话号码由专线号码/前缀或区域码和用户号码组成。区域或专线码识别用户电话所处国家内的区域，而用户号码是分配给特定专线码内用户的唯一识别号码。在世界许多地方电信服务提供商的撤销管制规定和私有化，为市场中多家新服务提供商的进入创建了条件。这使在一个给定区域内，通过拨打任何附加的运营商号码，用户能够从多家中选择服务提供商，成为必要的。用户可为本地、国家和国际呼叫，选择不同服务提供商。

交换和信令

电信网络的主要特征之一是一个独立的或带外信令网络，用来建立端到端呼叫。实际的呼叫（话音或数据）则通过一个独立的话音或数据网络传输。正常情况下，在电信网络中这些交换中心被称作交换局。用户拨打一个目的号码，则为信令网络建立到目的节点的一个呼叫提供了所需要的信息。信令网络端点传输控制信息，从一端到另一端建立电路。

电话交换局的主要任务是在两个用户端之间建立一条物理电路交换连接。但是，最近时间以来，多数交换局已经数字化，这包括端到端连接，由此将话音呼叫基本上变为数据呼叫，例外情况是，从交换局到用户连接的最后一段（还是电路的）。交换局一定是一个计算机系统，由外设实施电路交换和分组交换连接。

早期，信令信息用来从一个交换局传递到另一个交换局，其中使用与信令相关联的信道（CAS），这是一种带内信令机制。通用信道信令（CCS）已经替代了早期的CAS信令。在CCS中，SS7协议被用来交换信令帧，建立一条连接。SS7信令网络组件如图3.37所示。

以一种层次结构拓扑互联各信令交换局。这在便利网络管理方面提供帮助，并使一个呼叫的网络路由变得相对直接。如果目的用户没有位于由交换局所服务的区域，则每个交换局将呼叫沿层次结构向上路由。每个国家的电话网络的层次结构拓扑，随国家而不同。

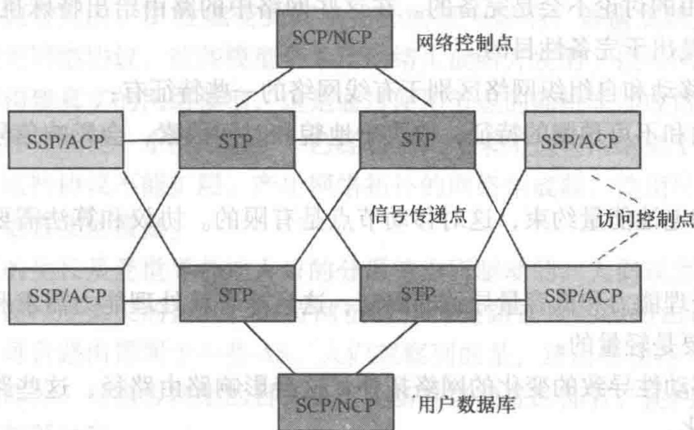


图 3.37　AIN（高级智能网）SS7 信令网络组件

从所接收到的信令信息中，交换系统确定地址信息，确定到达或去往目的地的路由，之后依据需要改变代码［例如自动交错路由（Automatic Alternate Routing），如在紧急号码的情形］之后进一步推进代码（接近目的地）。基本路由的隐喻（Metaphor）是层次结构的。如果目的地呼叫没有对应于一个用户站，即不在由当前交换局寻址的区域内，则该呼叫被路由到一个上级（Upward）交换局，将之朝目的地路由。

每个国家至少有一个国际交换中心，各专线（Trunk）交换局都连接到它。通过这个最高的交换层次结构级，从一个国家到另一个国家连接国际呼叫，任何用户能够通达世界上任何其他用户。在现代交换系统中，针对大体量的流量，呼叫可被直接连接到另一个低层交换机。

本地电话交换局可分析整个电话号码，旁路交换层次结构，并直接路由呼叫，条件是目的地是一个邻居本地交换局的一名用户。在 IN（智能网）的情形中，被拨打的号码是一个逻辑号码，可连接到一个物理用户号码，这取决于某些条件。

国家交换网络是层次结构的，在本地交换局之上包含多个交换机层次。本地交换局被连接到专线交换局，后者被连接到其他更高层的专线交换局，接下来可能被连接到同一服务提供商或另一个服务提供商的其他专线交换局，如在国际交换局的情形。注意，这里与各 ISP 如何连接到另一个 ISP 具有相似性，至少在层次结构的较高层次是这样的。通过高容量传输路径，专线交换局被连接到另一个专线交换局。这些专线总是有替代路径。这被称作一个骨干网络。

国际交换局可通过水下海底电缆、卫星链路或微波无线电链路进行连接。

3.11　无线、移动、自组织和传感器网络中的路由

在没有至少给出无线、移动和自组织以及传感器网络中路由的简短描述条件

下，对网络路由的讨论不会是完备的。在这些网络中的路由给出特殊挑战和问题，在这里提一下是出于完备性目的。

将无线、移动和自组织网络区别于有线网络的一些特征有：

1）可变的和不可预测的特征，取决于地貌和时间因素，会影响信号强度和时延特征。

2）带宽和电池能量约束，这对移动节点是有限的。协议和算法需要将这些因素考虑在内。

3）由于处理能力和低容量导致的限制，这意味着就处理能力需求和存储要求方面，协议需要是轻量的。

4）由于移动性导致的变化的网络拓扑，这会影响路由路径，这些路径会保持频繁地发生变化。

无线媒介是一种广播媒介，这意味着传输范围内的所有节点可听到广播，所以链路层协议需要能够处理冲突。

就计算和存储需要来说，适用于有线网络的路由协议具有高的开销，这使它们不适合于无线网络。结果，用于移动和自组织网络的多数路由协议作为有线网络路由协议的高度优化版本进行演进，或独立地针对移动和自组织网络而进行设计。有关这个专题的整个讨论是极尽穷举的，并在本书其他地方讲解。

3.12 网络、复杂性的本质和其他创新

因特网是一个网络。路由器是网络的一个至关重要的单元，在网络内提供信息路由服务。但准确地说，什么是网络呢？

将退一步而不是退几步来讨论，并研究在各学科中网络的角色，考察一下正在形成的网络科学。几乎在所有学科中均可找到网络，如基因网络、因特网和万维网，它们被定义为具有复杂拓扑的网络。不考虑基础现象，所有复杂网络共享某些共同的性质和机制。其中两个可描述为：通过添加新的节点，网络具有不断扩展的倾向；新节点首选地附接到其他节点，后者已经是良好连通的（集群倾向）。

如在许多其他学科中观察到的，大型网络的发展刻画出一种基础自组织倾向，上述两种机制是这种倾向的一个直接结果。

在这样一个网络中的节点分布被称作尺度无关（Scale-Free）分布。

早期，这些网络被建模为随机网络，是由 Paul Erdis 的重要工作假定得到的。但是，最新的研究表明，许多这样的网络是尺度无关分布的，而这些是背离于随机网络的。如今这些网络被看作演变的动态系统，具有清晰的基础规律，而不是静态的或随机图。

通过增加新节点和节点之间的链路，这些网络得以持续增长。在这样一个网络上任意节点 N 处获取新节点的速率，符合一个幂律函数而不是一个线性函数。

如在前面看到的，IP 性能受到网络拓扑的极大影响。普遍实践是，在网络模型上测试新的网络协议，这些模型是采用网络生成器开发的。这些网络生成器中的多数生成器构建真实的网络模型，这是通过使用节点的随机添加或指数添加派生得到的，以此种方式使一个网络生长。已经看到，当采用这些网络模型测试时，对于大型网络，这些协议不能扩展。产生网络拓扑的网络生成器，使用尺度无关的生长范型，产生更真实的模型。

因特网的生长是受世界各地人口的分形特点所驱动的。人们观察到，因特网的增长模式遵循尺度无关的拓扑。因特网被看作由链路连接的各路由器（节点）的一个网络，每台路由器属于一些 AS。人们观察到的是，路由器和各 AS 是一个分形集合的组成部分。可编程网络的目标是简化新网络服务的部署，使网络显性地支持服务生成和部署过程。

在这里值得提到的一个新的技术方向，是可编程网络。

在网络领域中发生许多创新，这是由于多种趋势和技术的融合产生的。一个这样的范型改变模型是网络可编程能力模型（见图 3.38）。这涉及网络硬件和控制软件、编程网络接口的开放 API、网络基础设施的虚拟化等的分离。这导致可定制和新的网络服务以及环境的快速开发和部署的趋势，即提供各种 QoS 流量整形模型。这被称作可编程网络。可采用开放可编程网络 API 和各种服务组合工具箱做到这一点。

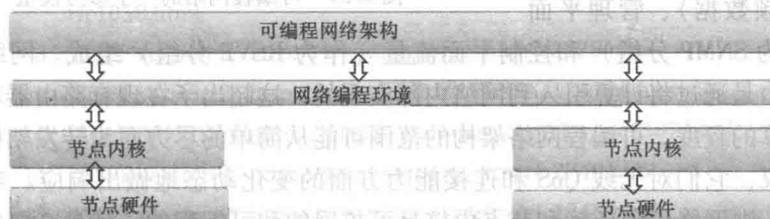


图 3.38　一种可编程网络架构的逻辑模型

当前，在路由和交换硬件与运行在其上的软件之间提供一定程度的隔离是非常困难的，这是由于它们之间的常见紧密耦合和相互依赖关系导致的。正常情况下，即使在今天，服务提供商对专用硬件和由路由器制造商开发的软件或称之为节点的（设备）也不能访问。

通过控制节点硬件的状态，节点内核被提供节点状态的可编程能力。节点内核使用户可共享节点的计算资源和通信资源。节点硬件可以是一台 IP 路由器或交换机、一台基站或一台媒体网关。由节点内核提供的低层接口被用作一个子层，构建更复杂的网络层服务。

存在一些工具箱，提供它们是为了开发和部署可编程网络。多数这样的网络遵循两个哲学理念之一，是由主动网络（AN）或开放信令（Opensig）共同体所支持

的。一个网络编程环境为构造网络架构提供一组网络编程接口。从哲学角度来说，这类似于使用软件开发工具箱构造新应用。但是，在这种情形中，应用是网络架构。

Opensig共同体采取从电信网络模型派生出的一种方法，其中通信硬件提供一组开发接口API，它们支持第三方硬件和软件提供商为附加服务提供插件或附加件，这样就允许较新架构和服务的就绪提供。一个可编程网络被逻辑上看作由一组不同的层或模块组成，如传输、控制和管理。这里，重点是服务生成。在这种情形中，IP交换机、路由器和移动网络被建模为带有良好定义的开放接口的分布式计算对象。这些对象的服务可使用通用对象请求代理架构（CORBA）或基于远程方法触发（RMI）的中间件工具箱加以访问。

图3.39所示为一个参考模型，它将因特网参考模型与控制、传输和管理平面组合在一起。在路由器的情形中，作为个体节点，在前面的讨论中也看到了这个模型。但是，这里这个模型被扩展为可编程网络架构。如果要刻画通过网络的多媒体流化，则它由传输平面流量（作为UDP分组携带视频数据）、管理平面流量（作为SNMP分组）和控制平面流量（作为RSVP分组）组成。网络服务的可编程能力是通过将计算引入到网络内部实现的，这超出了在现有路由器和交换机中实施计算的程度。可编程网络架构的范围可能从简单的尽力而为转发架构到复杂的移动协议，它们对无线QoS和连接能力方面的变化动态地做出响应。考虑到这种多样性，则网络编程环境和节点内核是可扩展的和可编程的，就是必要的，以便支持大量种类的可编程网络架构。

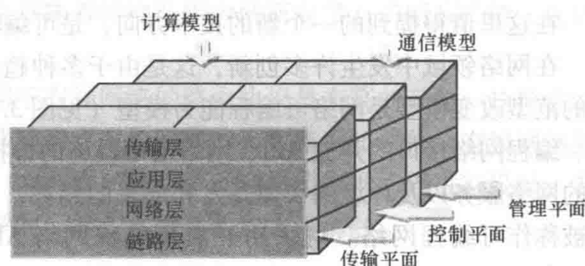


图3.39　可编程网络的一个参考模型

计算平面提供了这样一种能力，即在由路由器/交换机/节点组成的结构上、在网络中、传输/控制和管理平面间开发可编程服务的能力。

主动网络共同体支持“主动分组”或“密封小容器”（Capsule）的思路，它被用于在运行时动态地部署服务。在一些情形中，这些“密封小容器”由可执行程序或Java代码和数据组成。在主动网络中，代码移动性为部署新服务提供了主要载体。由此，网络行为由传输到网络上各节点的代码使用所修改，而不是通过可编程控制平面使用准静态网络编程接口实现。虽然这个范型为主动网络的部署提供了最大的灵活性，它也向编程模型提供了更多的复杂性。

正在进行的各种项目使用两种方法——探索可编程网络架构和服务的方法。一种这样的试验型设置，称作智能分组，基于主动节点方法，是在美国堪萨斯大学开发的，作为一个可编程IP环境实现的。基本上而言，智能分组是使用Java类构建

的移动代码，以带内和带外信息的方式传播。部署这种服务的基础设施由资源控制器、节点管理器和状态管理器组成。资源控制器提供到节点资源的接口。节点管理器控制资源利用率，而状态管理器管理各节点的状态，也管理整个系统的状态。它也集成了一种反馈调度算法，这支持资源使用情况的监测，方法是监测主动网络实体。

参考文献

1. Dave Roberts, *Internet Protocols Handbook*, Coriolis Group.
2. Daniel Lynch and Marshall Rose, *Internet System Handbook*, Addison-Wesley.
3. Douglas Comer and David Stevens, *Internetworking with TCP/IP, Volume I—Principles, Protocols and Architecture* (3rd edition), Prentice Hall.
4. Douglas Comer and David Stevens, *Internetworking with TCP/IP, Volume I—Design, Implementation and Internals*, Prentice Hall.
5. Colin Smythe, *Internetworking—Designing the Right Architectures*, Addison-Wesley.
6. http://www.faqs.org/rfcs/
7. "Routing information protocol," RFC 1058, http://www.faqs.org/rfcs/rfc1058.html.
8. "OSPF version 2," http://www.faqs.org/rfcs/rfc2178.html.
9. "OSPF version 2," http://www.faqs.org/rfcs/rfc1247.html.
10. "The OSPF NSSA option," http://www.faqs.org/rfcs/rfc1587.html.
11. "OSPF version 2," http://www.faqs.org/rfcs/rfc2740.html.
12. "OSPF version 2," http://www.faqs.org/rfcs/rfc1245.html.
13. "BGP OSPF interaction," RFC 1403, http://www.faqs.org/rfcs/rfc1403.html.
14. "Multicast extensions to OSPF," RFC 1584, http://www.faqs.org/rfcs/rfc1584.html.
15. "Extensions to OSPF to support mobile ad hoc networking," RFC 5820, http://www.faqs.org/rfcs/rfc5820.html.
16. "A border gateway protocol 4 (BGP-4)," RFC 1771, http://www.faqs.org/rfcs/rfc1771.html.
17. "BGP-4 protocol analysis," RFC 1774, http://www.faqs.org/rfcs/rfc1774.html.
18. Uyless D. Black, *IP Routing Protocols: RIP, OSPF, BGP, PNNI, and Cisco Routing Protocols*.

19. Deepankar Medhi and Karthikeyan Ramasamy, *Network Routing: Algorithms, Protocols, and Architectures.*
20. John T. Moyi, *OSPF: Anatomy of an Internet Routing Protocol.*
21. Jeff Doyle and Jennifer DeHaven Carroll, *Routing TCP/IP.*
22. William J. Dally and Brian Towles, *Principles and Practices of Interconnection Networks.*
23. Brian M. Edwards, Leonard A. Giuliano, and Brian R. Wright, *Interdomain Multicast Routing: Practical Juniper Networks and Cisco Systems.*
24. "VRRP: increasing reliability and failover with the virtual router redundancy protocol."
25. Prasant Mohapatra and Srikanth Krishnamurthy, *Ad Hoc Networks: Technologies and Protocols.*
26. C. S. Raghavendra, Krishna M. Sivalingam, and Taieb Znati, *Wireless Sensor Networks.*
27. Kazem Sohraby, Daniel Minoli, and Taieb F. Znati, *Wireless Sensor Networks: Technology, Protocols, and Applications.*
28. Morgan Kaufmann, *Computer Networks: A Systems Approach, 3rd Edition (The Morgan Kaufmann Series in Networking)*, 2003.
29. David Piscitello and Lyman Chapin, *Open Systems Networking: TCP/IP & OSI*, Addison-Wesley.
30. Buck Graham, *TCP/IP Addressing*, AP Professional.
31. Uyless Black, *TCP/IP and Related Protocols*, McGraw-Hill.
32. Martin Arick, *The TCP/IP Companion*, QED.
33. William L. Whipple and Sharla Riead, *TCP/IP for Internet Administrators*, EZine Publications.
34. Marshall Breeding, *TCP/IP for the Internet*, Meckler.
35. W. Richard Stevens, *TCP/IP Illustrated, Volume 1, The Protocols*, Addison-Wesley.
36. Sidnie Felt, *TCP/IP: Architecture, Protocols and Implementation*, McGraw-Hill.
37. K. Washburn and J. T. Evans, *TCP/IP: Running a Successful Network*, Addison-Wesley.
38. S. Floyd and V. Jacobson, *The Synchronization of Periodic Routing Messages*, April 1994.
39. Internet Assigned Numbers Authority, "Port numbers," Plain text, May 22, 2008, http://www.iana.org/assignments/port-numbers. Retrieved 2008-05-25.
40. C. Hendrik, *RFC 1058, Routing Information Protocol*, Internet Society, June 1988.

41. G. Malkin, *RFC 1388, RIP Version 2—Carrying Additional Information*, Internet Society, January 1993.
42. G. Malkin, *RFC 2453, RIP Version 2*, Internet Society, November 1998.
43. R. Atkinson and M. Fanto, *RFC 4822, RIPv2 Cryptographic Authentication*, Internet Society, January 2007.
44. "Implementation of a sensor network," http://today.cs.berkeley.edu/800demo/.
45. Alex Galis, Spyros Denazis, Celestin Brou, and Cornel Klein, *Programmable Networks for IP Service Deployment*.

第4章 全IP网络：移动性和安全性

Asoke K. Taluker

4.1 引言

移动性为受限在一定范围带来了自由。受这种自由思想的驱动，如何使消费者不受羁绊，一直是业界和研究人员的关注领域。在一个世纪以前，无线电报就使用没有导线的通信方式。在后来的阶段，话音也变为无线的。自1979年以来，无线数据就存在了[1,2]。但是，挑战是如何实施一项工业等级的通信无线服务，其中数千人可使用这种服务，而且，毕竟用户会在世界各地到处移动，并在没有任何约束的条件下使用无线服务。使用漫游技术，全球移动通信系统（GSM）技术解决了话音通信的这项挑战。

移动性要求具备如下特征：

1）无导线的通信。

2）感知到用户在哪里或可能在哪里（对于到达呼叫）。

3）如果用户位置未知，则具有如何通过寻呼定位用户的知识。

4）在一个网络内和网络间呼叫路由和连接的管理。

5）用户认证和交换安全密钥，确保通信是得到安全保障的，且由异地网络提供的服务得到支付。

通过无线电技术实现了第1点。但是，这需要增加无线频谱高效使用和以最小能量发送无线电波的技术支持。虽然看起来是简单的，但第2、3、4和5点是非常复杂的。GSM解决这些挑战的方法是，使用连接在信令网络中的数据库，所有服务提供商均可访问这个数据库。在GSM中使用的有两个数据库，即归属网络（HN）中的归属位置寄存器（HLR）[60]和拜访或服务网络（VN）中的拜访者位置寄存器（VLR）[61]。这些数据库是如此复杂，以致它们现在可提供号码便携能力[3,4]，即使用户已经改变HN并携带相同的标识符（电话号码）到另一个网络，也能定位用户。

GSM和话音中国际漫游的成功，激励因特网协议（IP）共同体也考虑移动性（问题）。在1996年，通过标题为“IP移动性支持”的请求评述（RFC）2002[8]做出了第一次尝试。IP域中移动性的基本思想基于GSM的类似哲学理念，即通过两台路由器（带有数据库和路由表）管理复杂性，即本地代理（HA）和外地代理（FA），其中HA提供与HLR的类似功能，FA具有像VLR的类似功能。采用许多

次修订，IP移动性进行了多次更新。最近的一次修订是2010年11月这么晚才做出的RFC 5944[59]，标题为“对IPv4的IP移动性支持，修订版”。

本章将讨论IP版本4（IPv4）和IP版本6（IPv6）的IP移动性，也讨论IPv6与一般而言IP中的漫游和切换。当谈论移动性时，移动节点（MN）将处在HN外部，并容易遇到安全威胁。因此，就移动性而言，也讨论IPv6中的安全性。

4.2 移动IP

移动IP背后的动机是，为一名用户提供一个环境，其中用户将在移动状态中在IP之上能够连续地访问数据和多媒体服务。在常规IP中，当一名用户从一个子网移动到另一个子网时，附接点将改变，这将强制连接中止。实际上，在改变它到因特网的链路层附接点之后，一个MN必须能够与其他节点通信。支持这一点的技术是移动IP。移动IP的主要RFC有RFC 2002、RFC 2003、RFC 2004、RFC 2005、RFC 2006、RFC 3220和RFC 5944[8-12,26,59]。

通过传输控制协议/因特网协议（TCP/IP）网络的两个端点之间的一条数据连接，要求源端处的一个源IP地址和一个TCP端口，及其对端，带有目的IP地址的一个等价目的地TCP端口。与TCP端口组合使用的IP地址构成一个端点的一个附接点，源和目的端点形成一条连接。本质上而言，所有这四个实体（四元组）保持不变（物理的或虚拟的），确保无缝通信。在移动状态中，一个MN的附接点将改变并中断连接。为修正这个问题，移动IP规范使MN能够使用由HA和FA分配的两个IP地址。这些IP地址分别被称作本地地址和转交地址。

一个MN被赋予一个HN上的一个长期IP地址（本地地址）。这个本地地址以一个“永久”IP地址被提供给一台静态主机的方式加以管理。在移动IP的语境中，一个MN或一个移动代理可被定义为一台主机或路由器，它可从一个网络（或子网）到另一个网络改变它的附接点。当远离其HN时，一个“转交地址”与MN关联，并反映MN的当前附接点。MN使用本地地址作为它所发送所有IP数据报的源地址。HA被定义为一个MN的HN上的一台路由器，它维护MN的当前位置信息，并当MN远离其本地时以隧道方式将数据报交付给MN。相对而言，一个FA被定义为MN的拜访网络上的一台路由器，它为MN提供路由服务。FA对由MN的HA以隧道方式传输的数据报去除隧道，并将之交付给MN。对于由一个MN发送的数据报，FA可作为注册MN的一台默认路由器。简单地说，到一台MN的所有到达分组通过一台HA被路由；对一条外发分组，则是不必要的。HA是MN的HN上的一台路由器，当MN远离本地时，HA为将数据报通过一条隧道交付给MN，而转发这些数据报。

当一台MN处在其HN中时，它在没有移动服务的情况下工作。当MN移动到一个外地网络时，它检测到它已经移动到一个外地网络，方法是将其自己的网络地

址（本地地址的最高24比特）与附接网络路由器的网络地址加以比较。它注册到FA，并从外地网络得到一个转交地址。转交地址可以是从一个FA的公告中确定的或采用某种外部指派机制，如动态主机配置协议（DHCP），见RFC 2131[13]中的解释。为使路由无缝地发生，HA需要知道MN的位置。因此，MN将其新的转交地址注册到其HA，通知它的新位置和新的转交地址。

举一个例子，有两个节点A和B，如图4.1所示。在这个例子中，节点A是静态的，节点B是移动的。当节点A发送一条分组到MN（例子中的节点B）时，则它将分组发送到节点B的本地地址，原因是节点A不知道节点B已经移出HN，且目前注册到另一个网络。举这样一个例子，考察一下移动IP数据报是如何在节点A和节点B的一条TCP连接之上交换的。

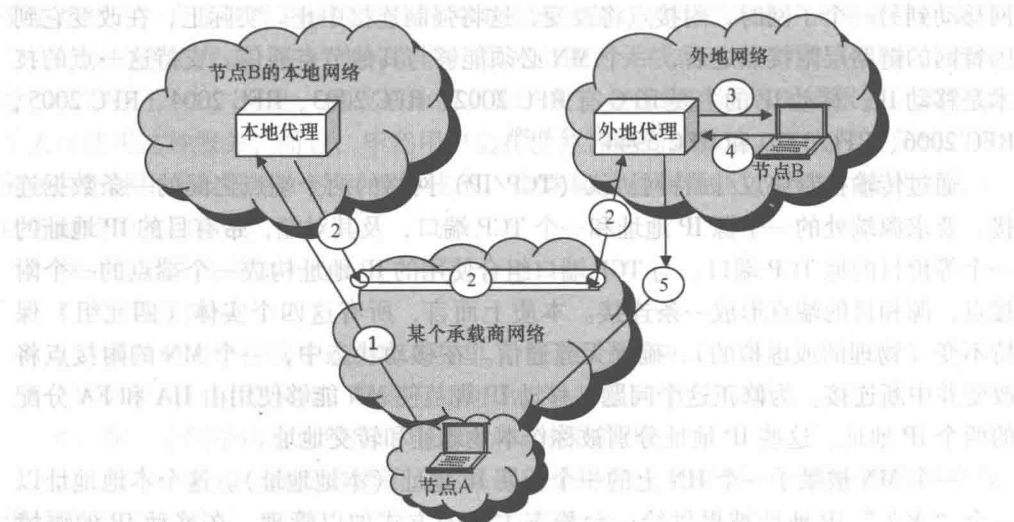


图4.1 移动IP架构

1）节点A希望将一条IP数据报发送到节点B。通告节点B的本地地址，则节点A知道这个地址。节点A不知道节点B是在HN还是在其他某处。因此，节点A将分组发送到节点B，以节点B的本地地址作为IP首部中的目的IP地址。该IP数据报被路由到节点B的HN。

2）在节点B的HN，到达IP数据报由HA截获。HA发现节点B处在一个外地网络中。这个外地网络将一个转交地址分配给节点B，且HA知道这个地址。HA将整个数据报封装在一条新的IP数据报内部，在IP首部中带有节点B的转交地址。HA重传这条新的数据报，转交地址为目的地址。在外地网络中，到达IP数据报由FA截获。FA是在外地网络中HA的对应实体。FA剥离外层IP首部，并将原数据报交付到节点B。

3）节点B拟对这条消息做出响应，并将流量发送到节点A。在这个例子中，

节点 A 是不移动的，由此节点 A 有一个固定 IP 地址。节点 B 使用节点 A 的 IP 地址作为 IP 首部中的目的地址。

4）由节点 B 到节点 A 的 IP 数据报使用节点 A 的 IP 地址作为目的地址，直接在网络间传输。因为节点 A 采用其本地地址，所以流量直接从节点 B 的转交地址到节点 A 的地址。

为支持上例中所示的操作，移动 IP 需要支持三项基本能力：

1）发现：一个 MN 使用发现规程，识别可加以使用的（有前景的）HA 和 FA。

2）注册：一个 MN 使用一个注册规程，将其当前转交地址通知它的 HA。

3）打隧道：使用一个打隧道规程，将 IP 数据报从一个本地地址转发到一个转交地址。

4.2.1 发现

代理发现是一个节点确定位置的方法。它确定它当前是连接到 HN 还是连接到一个外地网络。使用这个规程，一个 MN 可检测它何时从一个网络移动到了另一个网络。当连接到一个外地网络时，该方法也使 MN 能够确定 FA 转交地址是由那个网络上的每个 FA 提供的。代理公告消息是因特网控制消息协议（ICMP）路由器发现消息，由 HA 和 FA 传输，在一条链路上公告其服务。各 MN 使用这些公告来确定它们到因特网的当前附接点。移动 IP 发现规程是构建在一个现有 ICMP 路由器发现（路由器公告）之上的，路由器请求规程见 RFC 1256[7] 中 ICMP 路由器发现的规范描述。移动 IP 使用控制消息，这些消息发送到用户数据报协议（UDP）端口号 434，并从该端口发出。移动 IP 需要对当前消息格式做出扩展。通过 RFC 4066[37] 扩展了发现规程，以便支持 IP 切换。

在接收到这条公告分组时，MN 将路由器 IP 地址的网络部分与自己的 IP 地址（由 HN 分配的本地地址）的网络部分比较。如果这些网络部分不匹配，那么 MN 知道它处在一个外地网络中。一条路由器公告可携带有关默认路由器的信息和有关一个或多个转交地址的信息。如果一个 MN 需要一个转交地址而不等待代理公告，则 MN 可广播一条请求，这将由任意 FA 做出应答。

4.2.2 注册

一旦一个 MN 从外地网络得到一个转交地址，则同样需要注册到 HA。MN 发送一条注册请求到 HA，带有转交地址信息。当 HA 接收到这条请求时，它更新其路由表，并将一条注册应答发回 MN。

作为注册的组成部分，MN 需要加以认证。每个 MN、FA 和 HA 支持移动实体的一个移动安全关联（SA），由其安全参数索引（SPI）和 IP 地址索引指示该关联。在 MN 的情形中，这一定是它的本地地址。一个 MN 及其 HA 之间的注册消息

是采用一个支持授权的扩展加以认证的，如移动本地认证扩展。这个扩展是第一个认证扩展；在 MN 计算认证之后，将其他 FA 特定的扩展添加到该消息。使用一个 128 比特秘密密钥和基于哈希的消息认证码（HMAC）—消息摘要 5（MD5）哈希算法，产生一个数字签名。每个 MN 和 HA 共享一个共同的秘密。这个秘密使数字签名唯一，并使代理可认证该 MN。注册结束时，在 HA 中维护一个三元组，包含本地地址、转交地址和注册寿命。这被称作 MN 的一个绑定。直到注册寿命超期之前，HA 维护这个关联关系。注册过程涉及如下四个步骤：

1）通过向 FA 发送一条注册请求，MN 请求外地网络的转发服务。

2）FA 中继这条注册请求到那个 MN 的 HA。

3）HA 接受或拒绝该请求，并将一条注册应答发送给 FA。

4）FA 中继这条应答到 MN。

已经假定，FA 将分配转交地址。但是，可能的情况是，一个 MN 移动到没有 FA 的一个网络或这个网络上的所有 FA 都忙。也可能出现这种情况，转交地址由 MN 动态地获取一个临时地址，如 DHCP，见 RFC 2131[13] 中的解释，或由 MN 拥有，作为一个长期地址，仅在它访问某个外地网络时才使用。因此作为一种替代方法，通过使用一个共位转交地址，MN 作为其自己的 FA。一个共位转交地址是 MN 得到的一个 IP 地址，该地址与外地网络相关联。如果 MN 正在使用一个共位转交地址，那么直接注册到它的 HA。

4.2.3 打隧道

图 4.1 中的步骤 2 使用移动 IP 中的隧道传输操作。在移动 IP 中，使用 IP 内 IP 封装机制。在 IP 内 IP 中，HA 添加称作隧道首部的一个新 IP 首部。新隧道首部使用 MN 的转交地址作为隧道目的 IP 地址。隧道源 IP 地址是 HA 的 IP 地址。隧道首部使用 4 作为协议号，指明下一协议首部同样是一个 IP 首部。在 IP 内 IP 中，整个源 IP 首部被保留为隧道首部之净荷的第一部分。在接收到分组之后，FA 丢弃隧道首部，并将其他部分交付给 MN。

在任何 IP 数据分组中，源和目的 IP 地址必须是拓扑上正确的。移动 IP 中的前向隧道符合这条要求，因为其端点（HA 地址和转交地址）是其相应位置的合适指派的地址。另外，由 MN 发送的一个分组的源 IP 地址，没有对应于它所发出的网络前缀。为缓解这个风险，因特网工程任务组（IETF）提出反向隧道，是在 RFC 2344[14] 中规范的。

4.3 IPv6 的移动 IP

在 4.2 节中讨论了最初针对 IPv4 规范的移动 IP。在本节将讨论在 RFC 6275[60] 中规范的 IPv6 移动 IP（MIPv6），它包括许多附加特征。采用层次结构寻址方案的

IPv6，将能够相当高效地管理IP移动性。此外，IPv6尝试简化重新编址过程，这对因特网流量的未来路由能力是至关重要的。它保留了一个HN、一个HA以及使用封装将分组从HN交付到MN的当前附接点等思路。虽然仍要求发现一个转交地址，但一个MN可使用无状态地址自动配置和邻居发现来配置它的转交地址。由此，在IPv6中支持移动性，不要求有FA。

4.3.1　移动IPv6的基本操作

当一个MN处在HN中时，寻址到本地地址的分组被路由到MN的本地链路，使用的是常规因特网路由机制。当一个MN被附接到离开本地的某个外地链路时，它在一个或多个转交地址处也是可寻址的。MN可通过常规IPv6机制获取转交地址，这些机制如无状态或有状态自动配置。只要MN停留在这个位置，则寻址到这个转交地址的分组都将被路由到该MN。MN也可从几个转交地址接收分组，例如当它在移动过程中仍然在以前的链路上可达时的情况。

在IPv6的语境中，与一个MN通信的任何节点被称作MN的一个"通信节点"，该节点自己可以是一个静态节点或一个MN。对于MN和一个通信节点之间的通信，存在两种可能模式。第一种模式，双向隧道，不要求通信节点的MIPv6支持，即使MN没有将其与通信节点的当前绑定注册，也是可用的。来自通信节点的各分组被路由到HA，之后以隧道方式传递到MN。到通信节点的各分组以隧道方式从MN传递到HA（"反向隧道传递"），之后从HN以正常方式路由到通信节点。

4.3.2　移动IPv4和移动IPv6之间的差异

MIPv4和MIPv6之间的基本差异有：

1）不需要像在MIPv4中一样部署如FA的特殊路由器。在不要求本地路由器任何特殊支持的条件下，MIPv6可工作在任何位置。

2）对路由优化的支持，是协议的一项基本组成而不是一个非标准的扩展集。

3）即使在没有预先安排的SA（4.4节）情况下，MIPv6路由优化也可安全地操作。预期可在全球规模在所有MN和通信节点之间部署路由优化。

4）支持也被集成到MIPv6内，允许路由优化高效地与实施"进入过滤"的路由器并存。

5）IPv6邻居不可达检测，确保MN及其当前位置的默认路由器之间的对称可达性。

6）在MN离开本地时，在MIPv6中，发送到一个MN的多数分组，使用一个IPv6路由首部而不是IP封装发送，这样相比MIPv4，就降低了所产生的开销总量。

7）MIPv6与任何特定链路层解耦，原因是它使用IPv6邻居发现而不是地址解析协议（ARP）。这也改进了该协议的鲁棒性。

8）IPv6 封装（和路由首部）的使用，去除了在 MIPv6 中管理“隧道软状态”的需要。

9）MIPv6 中的动态 HA 地址发现机制将单条应答返回给 MN。在 IPv4 中使用的定向广播方法从每个 HA 处返回独立的应答。

4.3.3 移动 IPv6 安全

MIPv6 提供许多安全特征。这些特征包括对 HA 和通信节点之绑定更新的保护、移动前缀发现的保护和 MIPv6 用来传输数据分组之机制的保护。绑定更新是通过使用 IP 安全（IPsec）扩展首部或通过使用绑定授权数据选项加以保护的。这个选项利用一个绑定管理密钥 Kbm，可通过返回路由能力规程建立。通过使用 IPsec 扩展首部，保护移动前缀发现。与传输净荷分组相关的机制（如本地地址目的地选项和类型 2 路由首部）以一种约束其在攻击中使用的方式加以规范。

IPv6 中的 MN 和 HA 使用一个 IPsec SA，保护绑定更新和确认的完整性和真实性。MN 和 HA 支持和使用传输模式中的封装安全净荷（ESP）首部，并使用一个非空净荷认证算法，提供数据源发认证、无连接完整性和可选的防重放保护。

4.3.4 移动 IPv6 中的切换

在一个移动状态中，当一个呼叫在进行时，无线链路［接入路由器（AR）］和 MN 之间的关系是动态的。MN 可远离一个 AR 或接近它。当用户远离一个 AR 时，无线电信号强度或信号功率一直降低。这可导致连接中断。因此，为确保服务连续性，MN 必须将自己从以前的接入路由器（PAR）断开，并将自己连接到新的接入路由器（NAR），如图 4.2 所示。将链路从一个 AR 改变到另一个 AR 的这个规程被称作切换（Handover 或 Handoff）。在一系列 RFC 上讨论 IPv6 中的切换管理，即 RFC 4260、RFC 5268、RFC 5269、RFC 5270、RFC 5271 和 RFC 5380[39,50-53]。

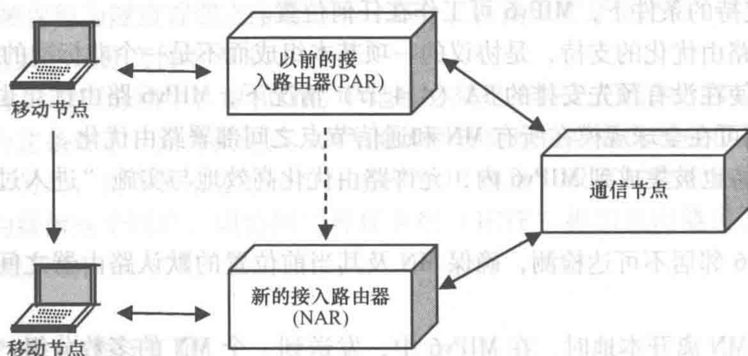


图 4.2 切换的参考场景

切换操作涉及链路层规程、运动检测、IP 地址配置和位置更新。在移动 IP 中，切换发生在两层中。当切换发生在链路层或无线接入点（AP）层时，这是一次 L2

切换。在一次L2切换之后，一个MN检测到一个在线子网前缀的改变，这将要求改变主转交地址。无线AP的改变典型地导致一次L3切换。

在一个IP网络中的切换是通过称作运动的一种哲学理念加以管理的。通用运动检测方法，使用邻居不可达检测技术，检测何时默认路由器不再是双向可达的。如果它是不可达的，则MN发现一台新的默认路由器（通常在一条新链路上）。运动检测的主要目标是检测L3切换。当MN检测到一次L3切换时，它在其链路本地地址上实施重复地址检测[23]，作为路由器发现的一个结果，选择一个新的默认路由器，之后采用那台新路由器实施前缀发现形成一个新的转交地址。之后它将其新的主转交地址注册到其HA。在更新本地注册之后，MN接着更新通信节点（它正在实施路由优化）中的关联移动绑定。

在相同链路上也许存在多台路由器，由此听到一台新的路由器，未必就构成一次L3切换。路由器的链路本地地址不是全局唯一的，所以在完成一次L3切换之后，MN也许继续以相同的链路本地源地址接收路由器公告。邻居不可达检测，确定默认路由器不再可达。对于一些类型的网络，发生一次L2切换的通知，也许是由低层协议或MN内的设备驱动软件得到的。一次L2切换指示，可能意味着L2运动切换或可能并不意味着L2运动切换，且L2运动可能意味着L3运动切换或可能并不意味着L3运动切换；相关关系也许是L2类型的一项功能，但也许是无线拓扑实际部署的一项功能。除非周知的情况下，一次L2切换指示极可能意味着L3运动切换，与直接组播一条路由器请求的做法不同，可能人们期望的是，验证默认路由器是否仍然是双向可达的。通过如下做法可做到这一点，发送一条单播邻居请求，并等待将被请求标志设置（为1）的一条邻居公告。在检测到它已经移动之后，一个MN使用正常的IPv6机制产生一个新的主转交地址。当前主转交地址过期时，也要完成这个过程。

当一个MN返回它的HN时，它通过运动检测算法，检测到它已经返回到它的本地链路。当MN检测到它的本地子网前缀再次上线时，完成检测。之后MN将一条绑定更新发送到它的HA，指令其HA不再截获或封装它的分组。通过RFC 2461[22]中描述的规程实施邻居发现。

4.3.5　3G CDMA网络之上移动IPv6中的切换

在RFC 5271[53]中描述了第三代（3G）码分多址（CDMA）网络的MIPv6快速切换。图4.3给出了支持移动IP的3G CDMA网络的一个简化参考模型。HA和MN的本地认证、授权和计费（HAAA）服务器驻留在本地IP网络，而MN在接入提供商网络内和之间漫游。通常，MN不处于本地IP网络中，相反，它们仅连接到接入提供商网络。在MIPv6注册之前，MN与AR建立一条3G CDMA接入技术特定链路层连接。当MN从一个AR移动到另一个AR时，重新建立链路层连接，实施一次MIPv6切换。那些AR驻留在同一个或不同的接入提供商网络中。在图

4.3 中，MN 通过无线电接入网络（RAN）从 PAR 移动到 NAR。在 3G CDMA 网络中，由基站（BS）发送的引导信道，使 MN 能够得到一个快速和准确的载波干扰比（C/I）估计。这个估计基于前向引导信道或引导强度的测量，这与一个 BS 的一个扇区相关联。

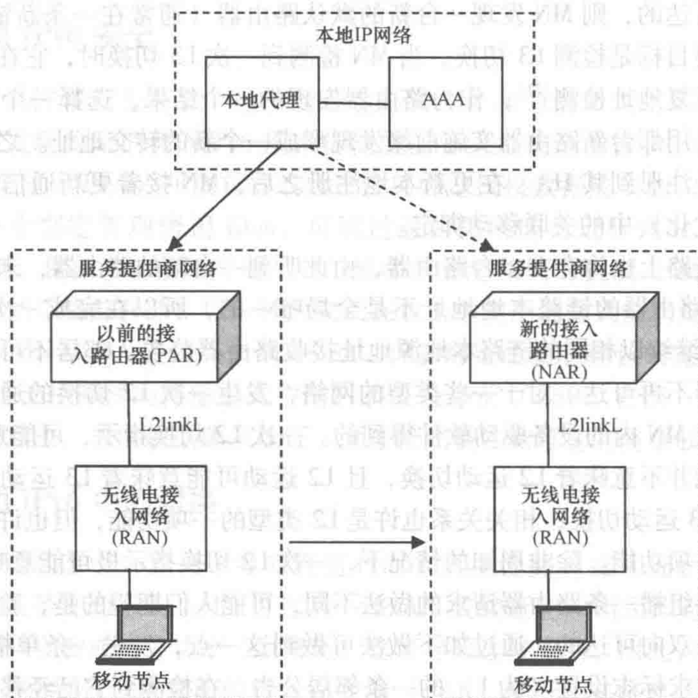


图 4.3　3G 网络上移动 IP 的参考模型

为辅助 MN 切换到新 AR，可考虑各种类型的信息：引导集，包括目标扇区或 BS 的候选者；MN 所在的蜂窝信息；RAN 中的服务节点；MN 的位置（如果存在的话）。为识别 MN 去往或离开的接入网络，可使用接入网络标识符（ANID）或子网信息[68,69]。在本文档中，这种新型的一个集合体被称作“切换辅助信息”。在 3G CDMA 网络中，在参考文献［50］中定义的新 AP 的链路层地址可能是得不到的。如果情况是这样的，则应该使用在这个文档中定义的切换辅助信息选项。

4.4 IP 网络中的安全

从电话商业化的那天起，电话公司就确保每个人要为他或她拨打的电话呼叫付费。不付费的任何人一定不被允许使用网络。因此，安全是电话网络的一个基本组成，在没有合适认证的条件下，不允许任何人进入网络。同样，在电信网络中存在大量智能（情报信息），这使实现漫游甚至号码便携能力变得容易。相反，当在 20

世纪70年代设计IP时，目标是提供简单的数据通信，简单性就是（那时的）咒语（Mantra）。例如，邮件传递的协议被称作简单邮件传递协议（SMTP），网络管理的协议被称作简单网络管理协议（SNMP）。简单是开放的，且容易调整；而安全性的基本要求则是受限的和得到控制的。同样，IP是为大学中的被信任用户设计的，且在IP网络中没有缴费和营账的需要。底线是安全性从来就不是IP网络的一项优先性任务。

如今，IP是数据和信息的一个主要载体，并必须保障安全。因此，挑战是使这样一个网络变得安全，而它在其核心中没有内建的安全原则。在IPv4和IPv6之间的最大区别之一是，期望所有的IPv6节点都实现强认证和加密特征，以便改进因特网安全性。IPv6原生地带有称作IPsec的一个安全协议，虽然许多厂商已经采纳IPsec，作为IPv4的组成部分。IPsec协议是为在IP网络间传递的信息提供机密性、完整性和真实性的一种基于标准的方法。IPsec将几项不同安全技术组合成一个完整的系统，提供安全性。

通过从协议集、密码学算法和密码学密钥中选择合适的安全属性，在网络层（IP层）提供IPsec安全服务。IPsec可被用来保护一条或多条“路径”：

1）一对主机之间。

2）一对安全网关（SG）之间。

3）一个SG和一台主机之间。

因为有时一台SG工作起来像一台主机，所以一台SG实现所有这三种形式的连接。一台遵循IPsec的主机也许不支持上述情况2，但必须支持情况1和3。

在主机实现中，IPsec可与操作系统集成。因为IPsec是一种网络层协议，所以它可作为网络层的组成部分加以实现，如图4.4所示。IPsec层需要IP层的服务才可构造IP首部。这个模型等同于其他网络层协议的实现，如ICMP。将IPsec与操作系统集成，有许多优势。一些关键优势如下：



图4.4 IPsec栈分层结构

1）因为IPsec是紧密集成到网络层的，所以它可提供网络服务，如分段、路径最大传输单元（PMTU）和用户语境（套接字）。这使实现非常高效。

2）因为密钥管理、基本IPsec协议和网络层可无缝地集成，所以按流提供安全服务（如一个web事务）是比较容易的。

3）支持所有的IPsec模式。

4.4.1 IPsec如何工作

以其基础功能引用IPsec架构的基础组件，如下描述：

1）SA：节点之间的关联（这些节点是什么，它们是如何工作的，它们是如何被管理的）和关联的处理，是在RFC 4301[40]中规范的。

2）安全协议：认证首部（AH）和ESP首部，分别在RFC 4302[41]和RFC 4303[43]中规范。

3）密码学算法：用于数据认证和加密的密码学算法，在RFC 4305[44]中规范。

4）密钥管理：任何安全协议都需要密钥来保障通信的安全。密钥管理是有关密钥生成［因特网密钥交换（IKE）］、存储和分发的（不管是人工的还是自动的），在RFC 4306中规范[45]。

SA的概念对IPsec而言是基础概念。在IP中，针对源和目的节点，使用一个IP地址和一个端口号，一起形成一个四元组，定义一条IP连接。类似地，IPsec使用一个SA跟踪有关两个节点之间一条给定IPsec连接的所有特定细节。一个SA是一个管理结构，用来为跨越IPsec边界的流量实施安全策略。一个SA如图4.5所示。SA处理的一项核心单元是基础安全策略数据库（SPD），它规定向IP数据报提供哪些服务，以及以哪种方式提供。

目的地址203.145.70.90
安全参数索引(SPI)937A1BC0
IPsec变换AH,HMAC-MD5
密钥A27574D2CFEA45A97E4F677329D84671
其他SA属性(未来)

图4.5 一个SA的例子

IPsec是一种点到点安全协议，其中一个SA维护端点（未必是端计算机）的所有安全相关的信息。它是一条逻辑的、单向的（单工）连接，为定义一条安全的电路，可定义为实体之间的关系。因为IPsec是一种点到点协议，所以这种安全关系包括主机、网关、防火墙甚至路由器，描述由IP地址和端口标识的端到端IP连接内的安全策略。如果希望保障两个支持IPsec的系统之间的双向通信，则需要两个SA，一个方向一个SA。

在一个实体或一个节点（不管它是一台计算机还是一个防火墙）中，将有许多条安全的IP连接，所以将有许多SA，它们被存储在一个安全关联数据库

(SAD) 中。为识别一个 SAD 内的一个特定 SA，必须有到数据库的一个指针，这个指针被称作一个 SPI。通过使用一个 AH 或一个 ESP（不能是两个都用），向一个 SA 提供安全服务。一个 AH 被用来提供完整性、数据源发认证和防重放攻击。相比而言，一个 ESP 提供机密性、完整性、真实性和防重放。如果 AH 和 ESP 保护被应用到一条流量数据流，那么必须创建两个 SA，并通过迭代地应用安全协议协同地实现保护。

每个 SA 由显式定义被保障安全的 IP 结构内一个点的一项安全特征的值组成，如目的地址、一个 SPI、那个会话使用的 IPsec 变换和其他属性，如 IPSec 寿命，如图 4.5 所示。

特别地，IPsec 使用如下密码学算法：

1）Diffie-Hellman 密钥交换机制，用来得到一个公众网络上两个实体之间的密钥。

2）公开密钥密码学，用来确保两个实体的身份，并避免中间人攻击。

3）块式对称密钥密码学，如高级加密标准（AES）、三重数据加密标准（3DES）等，用于数据的快速加密。

4）哈希算法（如 HMAC）与传统哈希算法［如 MD5 或安全哈希算法（SHA）］组合使用，提供分组完整性和认证。

5）数字证书，由一个证书权威签名，作为数字 ID 卡使用。

IPsec 使用多种密码学算法［用于块式数据（净荷）加密］和认证算法（用于 IPsec ESP 协议）。这些有三重 DES：加密块链式法（CBC）[21]、采用 128 比特密钥的 AES-CBC[31]、AES-CTR[32] 和 NOT DES-CBC，这在 RFC 2405[17] 中做了描述。对于哈希，它使用 HMAC-SHA1-96[10]、AES-XCBC-MAC-96[28] 和可选的 HMAC-MD5-96[15]。IPsec 的密钥交换协议类似于传输层安全或 TLS[49]。IPsec 因特网密钥交换版本 2（IKEv2）是在一系列标准中规范的，即 RFC 2407、RFC 2408 和 RFC 2409[19-21]。首部（HDR）包含 SPI、版本号和各种标志。SAi1 净荷声明发起者所支持的密码学算法。

4.4.2　IPsec 中的各元素

图 4.6 给出了总体 IPsec 架构。在这幅图中，“被保护的”（Protected）指位于 IPsec 保护边界内的系统或接口，而“不被保护的”（Unprotected）指位于 IPsec 保护边界外的系统或接口。被保护的接口可以是内部的，其中主机实现 IPsec，它可甚至连接到由主机操作系统给出的一个套接字层接口。一项 IPsec 实现工作在一台主机内，作为一个 SG 或作为一台独立设备，为 IP 流量提供保护。一个 SG 是实现 IPsec 的一个中间系统，可以是一个防火墙或激活 IPsec 的一台路由器。IPsec 提供的保护基于由一个 SPD 定义的需求，是由一名用户或系统管理员或由在其上面任意方建立的约束条件内工作的一项应用建立和维护的。一般而言，基于匹配 SPD

中表项的IP和下一层首部信息，针对三项处理操作之一选择各分组。每个分组使用IPsec安全服务被保护、丢弃或旁路IPsec保护，这取决于由选择器确定的可施用SPD策略。一个IPsec实现可支持边界任一侧或两侧上的一个以上的接口。

如前所述，在IPsec中，存在三个标称数据库：SPD、SAD和对端授权数据库(PAD)。

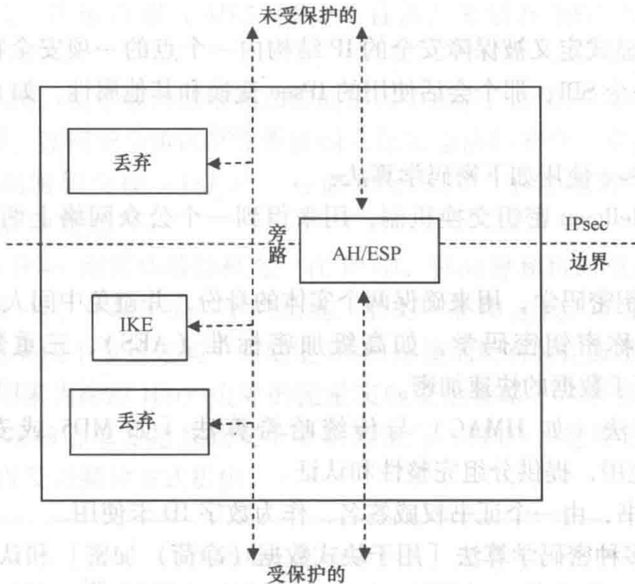


图4.6 顶层IPsec处理模型

SPD规范从一台主机或SG进入或外发的所有IP流量指派（Disposition）的策略。SAD包含与每个建立的［带密钥的］SA关联的各参数。第三个数据库PAD提供一个SA管理协议（如IKE）和SPD之间的一种联系。所有这三个数据库通过SPI中的表项联系起来。由IPsec提供的保护，基于由SPD定义的需求，SPD是由一名用户或系统管理员建立和维护的。当选择一项安全服务时，两个IPsec对端必须准确地确定要使用哪些算法。例如，AES-CBC用于加密，SHA-1用于完整性。在识别这项共同的使用需求中，IKE显式地创建SA对。PAD提供一个SA管理协议（如IKE）之间的一种联系。

SPD允许一名用户或管理员如下规范策略表项：

1）SPD-I：对于要被旁路或丢弃的进入流量，该表项由选择器的值组成，这些选择器应用到要被旁路或丢弃的流量。

2）SPD-O：对于要被旁路或丢弃的外发流量，该表项由选择器的值组成，这些选择器应用到要被旁路或丢弃的流量。

3）SPD-S：对于使用IPsec要加以保护的流量，该表项由选择器的值组成，这些选择器应用到通过一个AH或一个ESP保护的流量，基于这些选择器来控制如何

创建各SA，以及实施这项保护所需要的各项参数（如算法、模式等）。除了SPI外，一个SPD-S表项也包含这样的信息，诸如一个“从分组传递（Populate）”（PFP）标志与指明SA查找是否利用本地IP地址和远端IP地址的比特。

4.4.3　外发IP流量处理（保护到未保护）

术语“外发”指，通过保护接口进入实现的流量，或在边界的保护侧由实现发出的流量，被定向发往未保护的接口。当处理外发分组时，IPsec实施如下步骤：

1）当一条分组从用户（保护的）接口到达时，它触发SPD选择功能，得到选择合适SPD所需的SPD-ID。

2）针对由步骤1得到的SPD-ID指定的SPD，将分组首部与缓存进行匹配。

3）如果存在一个匹配，那么依据匹配缓存表项所指明的，使用一个AH或一个ESP，对分组进行处理，即旁路（BYPASS）、丢弃（DISCARD）或保护（PROTECT）。如果在缓存中没有找到匹配，则依据SPD-ID所指明的，搜索SPD（SPD-S和SPD-O部分）。如果SPD表项要求旁路或丢弃，则创建一个或多个新的外发SPD缓存表项，而如果旁路，则创建一个或多个新的进入SPD缓存表项。

4）分组被传递到外发转发功能，选择分组将被定向发往的接口。这项功能可能导致分组穿越IPsec边界被传回，进行附加的IPsec处理，如在支持嵌套SA的情况下。如果一个IPsec系统接收到必须被丢弃的一条外发分组，则它发送一条ICMP消息，向外发分组的发送者指明该分组被丢弃。

4.4.4　进入IP流量处理（未保护到保护）

术语“进入”（Inbound）指，通过未保护接口进入一个IPsec实现的流量，或在边界的未保护侧由实现发出的流量，被定向发往保护接口。进入处理不同于外发处理，原因是使用SPI将IPsec保护的流量映射到SA。进入SPD缓存（SPD-I）仅适用于旁路的或被丢弃的流量。如果一条到达分组看来是来自一个未保护接口的一个IPsec分段，则在IPsec处理之前要实施重组。在实施AH或ESP处理之前，通过未保护接口到达的任何IP分段都要重组（由IP实施）。将实施IPsec处理的每条进入IP数据报由AH或ESP值出现在IP下一协议字段中加以识别（或在IPv6语境中AH或ESP作为一个下一层协议出现）。

就一条进入分组，IPsec实施如下步骤：

1）当一条分组到达时，如有必要，为支持多SPD和关联的SPD-I缓存，它可能以所到达的接口（物理的或虚拟的）ID加以标记。接口ID被映射到一个相应的SPD-ID。

2）分组被检查，并解复用成两个分类之一：①如果分组看来是IPsec保护的，且它寻址到这台设备，则尝试通过SAD将之映射到一个活跃的SA。设备可有多个IP地址，可用于SAD查找，如在流控传输协议（SCTP）等各协议的情形。②流量

没有寻址到这台设备，或寻址到这台设备而不是一个 AH 或 ESP，则被定向实施 SPD-I 查找。如果利用多个 SPD，则在步骤 1 中指派给分组的标记被用来选择要被搜索的合适 SPD-I。SPD-I 查找确定动作是丢弃还是旁路。

3）如果分组寻址到 IPsec 设备，且一个 AH 或 ESP 作为协议被指定，则在 SAD 中查找该分组。对于单播流量，仅使用 SPI（或 SPI + 协议）。对于组播流量，使用 SPI + 目的地或 SPI + 目的地和源地址。如果不存在匹配，则流量被丢弃。如果分组没有寻址到设备，或被寻址到这台设备而不是一个 AH 或 ESP，则在（合适的）SPD-I 缓存中查找分组首部。如果存在一个匹配且分组将被丢弃或旁路，则执行这样的处理。如果没有缓存匹配，则在相应 SPD-I 中查找分组，并在合适的情况下，创建一个缓存表项。如果不存在匹配，则流量被丢弃。假定在 IPsec 边界的未保护侧发生 ICMP 消息的处理。

4）依据指定的情况，使用在步骤 3 中选择的 SAD 表项，实施 AH 或 ESP 处理。之后，将分组与由 SAD 表项识别的进入选择器匹配，验证所接收的分组对于通过它所接收的 SA 是合适的。

5）如果一个 IPsec 系统在一个 SA 上接收到一条进入分组，且分组的首部字段与 SA 的选择器不一致，则它丢弃该分组。为最小化一次拒绝服务（DoS）攻击或一个错误配置的对端的影响，IPsec 系统包括一项管理控制，允许一名管理员配置 IPsec 实现，发送或不发送这条 IKE 通知，而且如果选择这项设施，则要对这样的通知的发送实施速率限制。

4.5 融合网络中的认证、授权和计费

就处理安全和访问控制而言，IP 中和融合的 IP 多媒体系统（IMS）网络中，认证、授权和计费（AAA）协议所扮演的角色是极其清晰的。IMS 是一种架构性框架，在一个融合网络上交付 IP 多媒体服务，这种网络组合固定线路 IP 和无线 IP，提供数据和多媒体服务，服务范围从电子邮件到 IP 电视都有。它最初是由无线标准组织第三代伙伴项目（3GPP）设计的，作为将移动网络演进到 GSM 以远的愿景组成部分设计的（3GPP 规范组：R5，http：//www. 3gpp. org/ftp/Specs/html-info/TSG-WG—R5. htm）。3GPP、3GPP2 以及高等联网电信与因特网融合服务和协议（TISPAN）（http：//www. etsi. org/tispan/）更新了这项愿景。在前 IMS 时代，远程认证拨号用户服务（RADIUS）用于 AAA 服务。在 RFC 2865[24] 中定义了 RADIUS，在 RFC 2866[25] 中定义了 RADIUS 计费。在融合 IMS 中，Diameter 基本协议[29] 用于处理 AAA 功能。RADIUS 是一个缩略语，与此不同，Diameter（直径）不是一个缩略语，选择 Diameter 背后的原因是，它是 Radius（半径）的两倍。

4.5.1 Diameter

Diameter 基本协议是通过一组 RFC 定义的，即 RFC 3588、RFC 3589、RFC 4004、RFC 4005、RFC 4006、RFC 4072、RFC 4740、RFC 5224、RFC 5431、RFC 5447、RFC 5516 和 RFC 5624[19,30,34-36,38,47,48,55-58]，涵盖从 3GPP 到服务质量（QoS）的互操作标准。Diameter 基本协议优化 diameter 数据单元和能力的传递，以便协商和处理错误。Diameter 的主要性质有：

1）Diameter 是一个对等协议，意味着任何 Diameter 节点可发送或接收请求和应答到任何其他 Diameter 节点。

2）从基础传输层协议看，Diameter 期望由 TCP 提供的多数服务，如可靠性和拥塞控制。

3）Diameter 信令中的每个会话可包含几个个体请求和应答。

4）Diameter 将其节点分为三个不同类别：客户端、服务器和代理。客户端节点是在一个网络的边缘设备中实现的。服务器节点负责处理一个特定域的 AAA 请求。代理节点是提供中继、代理、重定向或转换服务的那些节点。

5）IMS 在许多接口中使用 Diameter。

6）Diameter 使用称作属性值对（AVP）的一种二进制首部格式和传输数据单元。

4.5.2 移动 IPv6 中的 AAA

RFC 5447[56] 描述 AAA 功能，这是在参考文献［46］中所述 MIPv6 启动解决方案所要求的，并将焦点放在网络接入服务器（NAS）到 HAAA 服务器通信的基于 Diameter 的 AAA 功能上。在集成场景中，作为网络接入认证规程的组成部分，提供 MIPv6 启动（方法）。图 4.7 给出了参与实体。其中，ASP 表示接入服务提供商；MSP 表示移动服务提供商；MSA 表示移动服务授权器。

在一个典型的 MIPv6 接入场景中，一个 MN 被附接到一个 ASP 的网络。在网络附接规程中，MN 与 NAS/Diameter 客户端交互通信。接下来，NAS/Diameter 客户端在 NAS 到 HAAA 接口之上与 Diameter 服务器交互通信。当 Diameter 服务器实施网络接入的认证和授权时，它也确定用户是否被授权使用 MIPv6 服务。在 MIPv6 服务授权和用户的策略概要基础上，Diameter 服务器可将几个 MIPv6 启动相关的参数返回给 NAS。在本书中描述的 NAS 到 HAAA 接口不与作为从 NAS/Diameter 客户端到 MN 传递 MIPv6 相关的配置参数唯一机制的 IPv6 DHCP（DHCPv6）绑定。

虽然这个规范解决 MIPv6 HA 信息的启动和可能的本地链路前缀指派问题，但它没有解决针对 MIPv6 目的如何在 MN 和 HA 之间创建 SA 的问题。

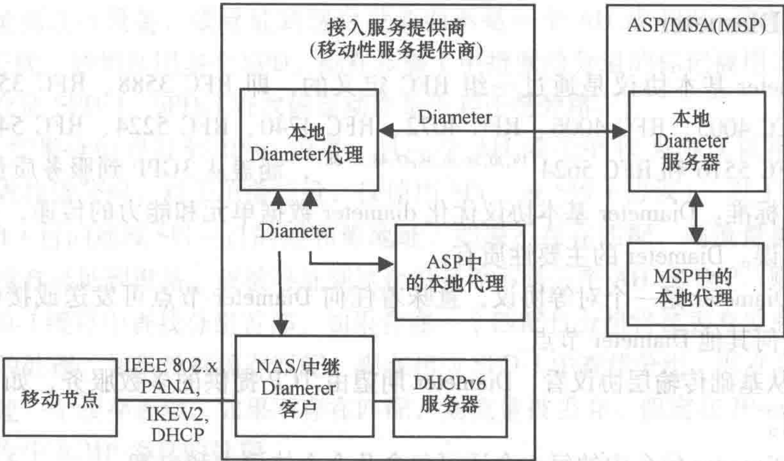


图4.7　集成场景中的MIPv6启动

4.5.3　一个融合的移动环境的安全框架

移动应用通常跨越几个网络。这些网络之一将是无线蜂窝无线电网络。其他网络将是有导线连接的网络。在任何这些网络的边界处，存在对协议转换网关的需要。这些网关运行在各层，包括应用层处的各个代理。多个网关和多个网络使移动环境中的安全挑战变得非常复杂。因此，为提供一个安全的移动环境，安全规程将是多个规程和功能的一个组合体。

4.5.4　3GPP安全

在由IP和电信组成的一个融合网络中，存在多项挑战。在一个计算机网络中的认证是在用户层完成的，而在电信中，网络认证是在设备层完成的。所有这些安全规则主要在于尝试保护运营商免于欺诈和网络滥用。

3GPP深入研究了这些担忧，并在当前无线广域网的安全架构中提出一些改变。3GPP通过重要的改变提出一种新的架构，如图4.8所示。

在下一代IMS融合网络中的安全考虑包括如下功能，像安全数据传输、认证、非否认（Nonrepudiation）、完整性、机密性、可用性、防重放和防欺骗。一个下一代融合网络将携带数据、运动图像和话音，其中话音将是电路交换（带有电话交换局的传统电信网）和分组交换［带有使用会话初始协议（SIP）的IP上的话音（VoIP，IP电话）］。在一个移动环境中，因为设备将是移动的，并从网络域到域地移动，就有必要从网络域安全（NDS）角度深入考察安全（问题）。在不同域和一个域的不同节点之间，提供IP安全能力，NDS可有所助益。在一个融合网络的语境中的一个安全域被定义为由单个行政管理权威运作的一个网络。在一个安全域内，为整个域维护一个统一的安全策略。一般而言，一个安全域将直接对应于核心

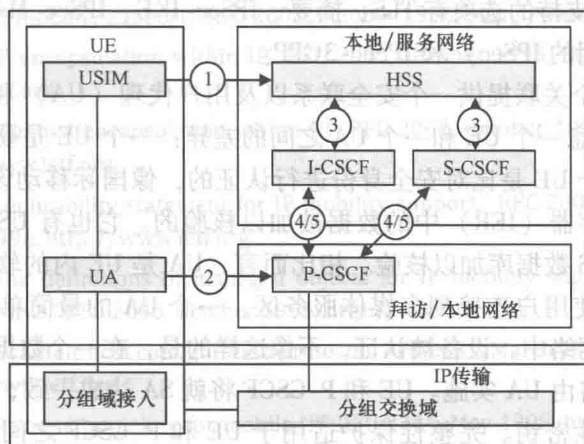


图4.8　IMS安全架构

网络。因此，一个融合IMS网络的安全考虑，需要解决域内和域间安全。此外，IMS需要解决接入安全和数据安全。针对安全，3GPP定义了如下标准：

1）安全架构和认证及密钥协议（AKA）[63]。

2）NDS[64]。

3）基于SIP服务的接入安全[65]。

4）通用认证架构[65]。

5）基于超文本传输协议（HTTP）服务的接入安全[67]。

当考察NGS安全时需要记住，它是计算机网络和电信网络的一个组合体。同样，在一个链中，最弱的连接才是链的强度。此外，将存在具有点到点和端到端安全的多个网关和代理。为提供这样复杂的安全需求（其中多媒体内容需要被交付到也许处于漫游状态的一名用户），需要以一种模块方式考察安全（问题），这是通过各种SA做到的。下一代网络（NGN）和IMS安全架构如图4.8所示。对于NGN的安全防护，存在5个不同的SA和不同需要，在图4.8中标号为1~5。

1）SA1：在这个关联中，实施用户设备（UE）和服务呼叫会话服务功能（S-CSCF）之间的双向认证。出于本章的考虑，CSCF被看作一组服务器或SIP代理。归属用户服务器（HSS）是下一代HLR，集体地负责AAA，关联的数据库将用户认证的实施委派给S-CSCF。HSS相应地负责产生安全密钥和加密与认证的挑战码。存储于统一用户身份模块（USIM）中的UE的长期密钥和HSS，是与用户的隐私身份相关的。USIM是下一代用户身份模块（SIM），它是一个智能卡安全设备，包含在一个移动网络中使用的用户的所有安全信息。用户将由一个（网络内部）用户隐私身份、国际移动隐私身份（IMPI）和至少一个外部用户公开身份、国际移动公开身份（IMPU）。UE和到运营商网络代理呼叫会话服务功能（P-CSCF）的第一个AP之间的SA，是在RFC 3329[27]中所定义协议的基础上协商

的。RFC 3329[29]支持的选项有TLS、摘要、IPSec-IKE、IPSec-MAN（没有IKE条件下采用人工密钥的IPSec）和IPSec-3GPP。

2）SA2：这个关联提供一个安全联系以及用户代理（UA）和一个P-CSCF之间的一个SA。注意一个UE和一个UA之间的差异：一个UE是硬件设备。在一个电话网络中，一个UE是比对安全身份进行认证的，像国际移动设备身份（IMEI）是与设备身份寄存器（IER）中的数据被加以核验的。它也有USIM，存储各种安全身份，对比HSS数据库加以核验。相比而言，UA是UE内的软件，其工作就像一个代理一样，使用户连接到多媒体服务区，一个UA的最简单例子是web浏览器。在一个电话网络中，设备被认证，不像这样的是，在一个数据网络中，用户被认证，这项功能将由UA实施。UE和P-CSCF将就SA达成一致，这包括要用于完整性保护的完整性密钥。完整性保护适用于UE和P-CSCF之间，用于保护所有通信。

3）SA3：这个关联以内部方式提供网络域内的安全。

4）SA4：这个关联提供不同网络之间的安全。当P-CSCF驻留在拜访网络（VN）内时，这个SA是可用的。如果P-CSCF驻留在HN中，则应用SA5。

5）SA5：这个关联以内部方式在IMS子系统内支持SIP的节点之间，提供网内安全。注意，当P-CSCF驻留在HN中时，也应用这个SA。

参考文献

1. Wikipedia, the free encyclopedia, http://www.wikipedia.org.
2. Worcester Polytechnic Institute, "The first IEEE workshop on wireless LANs: preface," http://www.cwins.wpi.edu/wlans91/scripts/preface.html. Retrieved March 16, 2008.
3. 3rd Generation Partnership Project, "Number portability, technical specification, 3rd generation partnership project; technical specification group services and system aspects; support of mobile number portability (mnp); service description; stage 1, (release 6), 3GPP TS 22.066 V6.1.0," June 2003, http://www.3gpp.org.
4. Asoke K. Talukder. "Mobile number portability: making SMS-data services portable," *Journal of Indian Institute of Science*, Mar.–Apr. 2006, **86**, 81–98.
5. Asoke K. Talukder, Hasan Ahmed, and Roopa Yavagal, *Mobile Computing, Technology, Application and Service Creation*, McGraw-Hill, 2010.
6. Asoke K. Talukder and Manish Chaitanya, *Architecting Secure Software Systems*, CRC Press, 2009.
7. "ICMP router discovery messages," RFC 1256, September 1991, http://www.ietf.org.

8. "IP mobility support," RFC 2002, October 1996, http://www.ietf.org.
9. "IP encapsulation within IP," RFC 2003, October 1996, http://www.ietf.org.
10. "Minimal encapsulation within IP," RFC 2004, October 1996, http://www.ietf.org.
11. "Applicability statement for IP mobility support," RFC 2005, October 1996, http://www.ietf.org.
12. "The definitions of managed objects for IP mobility support using SMIv2," RFC 2006, October 1996, http://www.ietf.org.
13. "Dynamic host configuration protocol," RFC 2131, March 1997, http://www.ietf.org.
14. "Reverse tunneling for mobile IP," RFC 2344, May 1998, http://www.ietf.org.
15. "The use of HMAC-MD5-96 within ESP and AH," RFC 2403, November 1998, http://www.ietf.org.
16. "The use of HMAC-SHA-1-96 within ESP and AH," RFC 2404, November 1998, http://www.ietf.org.
17. "The ESP DES-CBC cipher algorithm with explicit IV," RFC 2405, November 1998, http://www.ietf.org.
18. "IP encapsulating security payload (ESP)," RFC 2406, November 1998, http://www.ietf.org.
19. "The Internet IP security domain of interpretation for ISAKMP," RFC 2407, November 1998, http://www.ietf.org.
20. "Internet security association and key management protocol (ISAKMP)," RFC 2409, November 1998, http://www.ietf.org.
21. "The ESP CBC-mode cipher algorithms," RFC 2451, November 1998, http://www.ietf.org.
22. "Neighbor discovery for IP version 6 (IPv6)," RFC 2461, December 1998, http://www.ietf.org.
23. "IPv6 stateless address autoconfiguration," RFC 2462, December 1998, http://www.ietf.org.
24. "Management information base for IP version 6: textual conventions and general group," RFC 2865, December 1998, http://www.ietf.org.
25. "RADIUS accounting," RFC 2866, June 2000, http://www.ietf.org.
26. "IP mobility support for IPv4," RFC 3220, January 2002, http://www.ietf.org.
27. "Security mechanism agreement for the session initiation protocol (SIP)," RFC 3329, January 2003, http://www.ietf.org.
28. "The AES-XCBC-MAC-96 algorithm and its use with IPsec," RFC 3566, September 2003, http://www.ietf.org.

29. "Diameter base protocol," RFC 3588, September 2003, http://www.ietf.org.
30. "Diameter command codes for third generation partnership project (3GPP) release 5," RFC 3589, September 2003, http://www.ietf.org.
31. "The AES-CBC cipher algorithm and its use with IPsec," RFC 3602, September 2003, http://www.ietf.org.
32. "Using advanced encryption standard (AES) counter mode with IPsec encapsulating security payload (ESP)," RFC 3686, January 2004, http://www.ietf.org.
33. "Mobility support in IPv6," RFC 3775, June 2004, http://www.ietf.org.
34. "Diameter mobile IPv4 application," RFC 4004, August 2005, http://www.ietf.org.
35. "Diameter network access server application," RFC 4005, August 2005, http://www.ietf.org.
36. "Diameter credit-control application," RFC 4006, August 2005, http://www.ietf.org.
37. "Candidate access router discovery (CARD)," RFC 4066, July 2005, http://www.ietf.org.
38. "Diameter extensible authentication protocol (EAP) application," RFC 4072, August 2005, http://www.ietf.org.
39. "Mobile IPv6 fast handovers for 802.11 networks," RFC 4260, http://www.ietf.org.
40. "Security architecture for the Internet protocol," RFC 4301, December 2005, http://www.ietf.org.
41. "IP authentication header," RFC 4302, December 2005, http://www.ietf.org.
42. "IP encapsulating security payload (ESP)," RFC 4303, December 2005, http://www.ietf.org.
43. "Extended sequence number (ESN) addendum to IPsec domain of interpretation (DOI) for Internet security association and key management protocol (ISAKMP)," RFC 4304, December 2005, http://www.ietf.org.
44. "Cryptographic algorithm implementation requirements for encapsulating security payload (ESP) and authentication header (AH)," RFC 4305, December 2005, http://www.ietf.org.
45. "Internet key exchange (IKEv2) protocol," RFC 4306, December 2005, http://www.ietf.org.
46. "Problem statement for bootstrapping mobile IPv6 (MIPv6)," RFC 4640.
47. "Diameter session initiation protocol (SIP) application," RFC 4740, November 2006, http://www.ietf.org.

48. "Diameter policy processing application," RFC 5224, March 2008, http://www.ietf.org.
49. "The transport layer security (TLS) protocol version 1.2," RFC 5246, August 2008, http://www.ietf.org.
50. "Mobile IPv6 fast handovers," RFC 5268, June 2008, http://www.ietf.org.
51. "Distributing a symmetric fast mobile IPv6 (FMIPv6) handover key using SEcure neighbor discovery (SEND)," RFC 5269, June 2008, http://www.ietf.org.
52. "Mobile IPv6 fast handovers over IEEE 802.16e networks," RFC 5270, June 2008, http://www.ietf.org.
53. "Mobile IPv6 fast handovers for 3G CDMA networks," RFC 5271, June 2008, http://www.ietf.org.
54. "Hierarchical mobile IPv6 (HMIPv6) mobility management," RFC 5380, http://www.ietf.org.
55. "Diameter ITU-T Rw policy enforcement interface application," RFC 5431, March 2009, http://www.ietf.org.
56. "Diameter mobile IPv6: support for network access server to diameter server interaction," RFC 5447, February 2009, http://www.ietf.org.
57. "Diameter command code registration for the third generation partnership project (3GPP) evolved packet system (EPS)," RFC 5516, April 2009, http://www.ietf.org.
58. "Quality of service parameters for usage with diameter," RFC 5624, August 2009, http://www.ietf.org.
59. "IP mobility support for IPv4, revised," RFC 5944, November 2010, http://www.ietf.org/.
60. "Mobility support in IPv6," RFC 6275, http://www.ietf.org/.
61. "Home location register specification," GSM 11.31, http://www.etsi.org.
62. "Visitor location register specification," GSM 11.32, http://www.etsi.org.
63. "Security architecture and authentication and key agreement (AKA)," 3GPP TS 33.102, December 2002, http://www.3gpp.org.
64. "Network domain security (NDS)," 3GPP TS 33.310, September 2004, http://www.3gpp.org.
65. "Access security for SIP-based services," 3GPP TS 33.203, December 2009, http://www.3gpp.org.
66. "Generic authentication architecture," 3GPP TS 33.220, March 2008, http://www.3gpp.org.
67. "Access security for HTTP-based services," 3GPP TS 33.222, March

2006, http://www.3gpp.org.

68. "3GPP2 access network interfaces interoperability specification," 3GPP2 TSG-A, A.S0001-A v.2.0, June 2001.
69. "Interoperability specification for high rate packet 1 2 data (HRPD) access network interfaces—rev A.," 3GPP2 TSG-A, A.S0007-A v.2.0, May 2003.

第5章　转换扩展的家庭：步向基于IP的异构以用户为中心融合环境的下一步骤

Josu Bilbao，Igor Armendariz

本章描述家庭网络和扩展家庭场景的演进（包括水平的和垂直的传输环境），并将焦点放在用户基础设施适应于数字多媒体服务的变革方面，重点是全因特网协议（IP）架构。

多媒体内容输入数量的指数增长，以及在相同基础设施之上合并控制和多媒体需求的出现，要求一种融合场景。

为在数字的扩展家庭通信中得到所要求的融合，全IP战略将扮演一个关键角色。本章将描述一种融合愿景，焦点是IP网络开放系统互联（OSI）层和由异构网络分段组成的一项基础设施。

与本章有关的关键特征基于开发一种异构IP扩展家庭架构的协议和物理媒介融合，设想为家庭内和公寓、办公室、水平和垂直传输场景的扩展家庭。

下面各节将描述在家庭中通信演进的步骤，描述在未来家庭网络中步向新型IP服务的最适合架构方式中的技术里程碑和挑战。

5.1 引言

为位于家庭内的一名给定用户所实施的服务部署有一个相对简短的历史[6]。但是，从通信的角度看，就可用服务数量和要满足的需求而言，它已经发生了密集的演进过程。从模拟家庭变换为一个数字家庭，事实上是本章的讨论焦点。

在下一节，讨论家庭通信基础设施的历史演进过程和当前服务的需求。

之后，描绘了与创新型应用和高清（HD）服务相关的一个新场景。接下来，形象地说明了家庭骨干的隐性需求，重点突出技术方面的问题，这些使在整个家庭内服务分发所要求基础设施的部署以及一个给定位置的用户端与服务和应用提供商的互联成为可能。

在本章中关注的主要专题之一是为提供三重播放和四重播放服务，针对扩展家庭，提供一个全IP基础设施的愿景[28]。值得指出的是，三重播放指数据、音频和视频流的组合，而四重播放将移动性特征添加到前面提到的流中。

以这个目标为主线，将描述各种机制，这些机制提供组成家庭骨干的不同网络分段间的融合。虚拟基础设施的部署，即不需要任何走线安装的那些设施，被称作不需新导线（No-New-Wires），将被作为前述基础设施的骨干而加以强调。

一旦描述基础设施，则将分析不同场景和家庭骨干的应用领域。在本章中研究的关键问题之一是，在所描绘的基础设施上以一种泛在方式提供的以用户为中心的服务。在家庭网络处 IP 融合的基础上，详细研究了新的潜在电子健康（e-Health）服务的潜力。

之后，描述了一个扩展家庭的概念，包括与共管场景以及垂直和水平传输场景相关的不同网络分段的全 IP 基础设施中的集成。所以，家庭骨干的应用领域被扩展到任何位置，其中就像用户在家中一样，他或她可消费各项服务。

最后，本章以家庭骨干的主要未来挑战和最具影响的研究动向（就下一代基础设施的就续提供而言）的描述收尾。

以前

分析在 20 世纪 90 年代中家庭的最新状态，值得欣慰的是，存在不同的视听内容源，其中绝大多数都被当前系统所继承。在 20 世纪末，模拟无线电和电视（TV）服务特别地向前跨越了一步（突出出来）。在家庭中多数音频分发是以无线电为中心的，针对这个目的，使用了不同的调制标准，就像射频调制（FM）和幅度调制（AM）的情形那样。就 TV 而言，使用了不同的信号传输技术。这种做法的一个例子是在甚高频率（VHF）和超高频率（UHF）波段，使用模拟 TV 广播。此后，卫星广播技术使如下情况成为可能，其中在极长距离上传输视听内容，由此鼓励文化多元性并使新的商务模型建立成为可能。

所以，在 20 世纪期间在家庭中用于服务和内容接收的基础设施是基于一个天线的，它支持模拟调制的无线电服务被接收，另一个天线用于接收模拟 TV 和一个锅式天线用于通过卫星接收内容。从用于在家庭中接收视听内容的电信基础设施角度看，它可能突出的是，基础设施部署在公共住宅（也称作公寓）中。在这些类型的装置中，支持对特定数量的住户使用共享的接收器，由此削减了访问广播型视听内容所需的基础设施成本。

一旦信号由天线所接收，则用于从捕获期间恰好重新分发到再生设备的主要物理媒介，一直就是同轴电缆，使用模拟信号放大器，支持在家庭内的不同位置接收内容。除了还不存在术语“家庭网络”的事实外，我们可以说，这是第一代家庭网络。

用户端最感兴趣的功能之一是，能够将视听内容和数据存储于存储设备之中。出于这个目的，人们开发了具有有限功能的各种存储设备，但几乎没有任何接口可使它们在一个网络中互联起来。

因此，如在参考文献［5］中所述，UHF TV 一直被强调为杀手级应用，FM 无线电作为音频内容分发中的主导方式，而对多媒体内容的存储则有各种解决方案，这取决于应用领域：VHS 和 Betamax 成为占主导地位的视听存储媒介，而卡带首选用于音频，磁盘用于数据（以 51/4″和 31/2″软盘形式存储）。

上述的基础设施确实考虑了视听内容分发的下行链路连接（到达家庭方向）。但是，要强调的是，在20世纪期间，多数家庭有双向话音连接，即所谓的公众交换电信网络（PSTN）。这项基础设施被用来建立双向话音通信，电信运营商是主要的互联工具提供商。电话链路在家庭中变为无线链路，这多亏无绳手持电话数字增强无绳电信（DECT）标准。

在20世纪90年代末，我们时代的主要技术革命之一，是在网络互联的基础上开始的（就我们所知的角度看，诞生了因特网）。在不同远程用户的设备之间提供连接能力，使用电话基础设施，开启了一种新型服务图景和创新型应用的诞生。

家庭，过去作为电话布线的一个毛细血管端，所以在提供因特网接入的辅助下使用PSTN基础设施的事实，是新服务的快速繁荣和技术被端用户采用的原因。

最初，对于网页浏览服务以及使用电子邮件和即时消息通信服务，低带宽连接在过去是足够用的。出现了可能改变联网世界的成熟协议IP。这个协议在后来的几十年中成为融合协议。

数年来，通信基础设施过去受限于话音分发，但随着数字时代的繁荣和联网技术方面的改进，开启了一个新的服务时代。需求的指数性增长转变为家庭骨干概念的基础。

5.2　新的全IP家庭场景

如在前面所描述的，随着因特网的诞生和服务数字化，在家庭中的通信经历了一次极端的演化。

原则上来说，随着服务数字化时代的蓬勃发展，在整个家庭上要求数据流重新分发的应用数量发生了指数性增长。对于如下方面情况就是这样的，有多媒体内容服务、网络游戏部署、家庭自动化、智能报警系统等。

在基础设施革命中关键角色之一是休闲娱乐。特别地，视听内容重新分发是这个划时代变化的驱动车轮。下面描述了步向一个新的家庭场景这种变化中的最重要角色。

5.2.1　高清多媒体服务蓬勃发展

在21世纪第一个十年期间，一些最常用的媒介输入是继承自模拟时代的：

1）频率调制和幅度调制的模拟无线电（FM和AM）。

2）带有卫星接收的模拟TV（FM-TV）。

3）地面模拟TV（UHF和VHF）。

一些这样的服务有要被关闭的一个确定的最后期限，但仍然到处都是各种模拟媒体，绝大多数的人仍然使用它们接收视听内容。这意味着在数字世界和模拟世界之间存在一个转换点，共存是需要的。

新世纪伊始见证了新的数字输入的出现，这导致了分发多媒体内容之商务模型的修正，特别强调家庭中服务的可访问能力。这产生了一代内容分发的各种技术，是在无线电电子频谱的高效使用基础上产生的：

1）数字视频广播（DVB）：该组织称为DVB，被授权（Entrusted）进行数字视听内容广播的标准化，已经规范了不同标准，包括卫星传输（DVB-S）、在光纤和同轴线缆上的内容传输（DVB-C）和地面广播之上的多媒体服务传输（DVB-T）。制定这些标准是面向以具有极高质量和附加值的数字服务替换模拟服务的。DVB最新部署了第二代服务分发标准（DVB-S2、DVB-T2、DVB-C2等）。

2）数字无线电广播，注定要替换模拟无线电：数字音频广播（DAB）将替换FM模拟无线电，数字调频广播（Digital Radio Mondiale，DRM）将替换AM无线电。

3）因特网协议电视（IPTV）或以所要求的服务质量（QoS）在IP上分发多媒体内容：这包括因特网进入家庭，典型地是在xDSL连接上进入的，目的是重新分发多媒体服务和采用会话初始协议（SIP）的视频会议服务。

除了经典服务的数字化外，最近数年见证了将多媒体服务分发到家庭的不同频道的出现：

1）全球微波接入互操作性（WiMAX），作为农村地区和移动性的主要最后一英里技术。

2）手持设备的数字多媒体广播（DMB）：DMB和DVB-H（手持设备的DVB视频广播规范）。

3）宽带移动连接能力：统一移动电信系统（UMTS-3GPP）、高速下行链路分组接入（HSDPA）、高速上行链路分组接入（HSUPA）、长期演进方案（LTE）等。

因此，存在访问多媒体内容的多种方式，使用从模拟时代继承的无线电分发系统或使用新的基于IP的分发网络（见图5.1）。



图5.1 多媒体服务进入家庭

在内容世界中主要的新发展之一是出现了使用户交互性成为可能的服务，这要感谢来自家庭的一条回传信道（一条上行链路）。这是与社交网络服务交互和基于交互能力［如多媒体家庭平台（DVB-MHP）］上载多媒体内容或服务的情形，其中使用回传信道，一般基于IP网络，通过一台家庭网关或驻地网关实现的。

在已经描述的这个场景中，对存在一项基础设施的兴趣（将支持以一种融合方式在整个家庭上进行内容重新分发）是明显的。从内容存储角度看，音频、视频和数据已不再像模拟时代那样独立地加以处理。如今，用户自己使用相同的设备存储视频、照片、音频和数据。采用这种方式，像DVD、PVR、固态硬盘驱动、蓝光磁盘等存储设备的使用，突出了内容存储中融合的重要性。

现在的挑战是为用户提供具有如下功能的一个家庭网络，将支持内容源和存储设备与处在不同家庭位置的端用户互联。

5.2.2 通信流的重新分发

大概来说，多媒体内容构成服务包，对用户端利益而言具有最大直接反馈，所以，它被看作新的家庭网络基础设施部署的主要激励工具。但是，应该强调的是，除了提到的多媒体服务外，对用户端和内容生成方/提供商之间重新分发数据网络链路，存在逐渐增加的需要。从家庭任何位置进行的因特网访问，情况就是这样的。

如今，存在不同的移动连接机制，这些机制由到处移动的用户所用。尽管如此，一条固定连接的成本是相当廉价的，且所提供的带宽是相当高的。例如，非对称数字用户线路（ADSL）或ADSL2+以越来越具有竞争性的价格提供每秒数十兆比特（Mbit/s）的固定连接能力。

所以，这正在增加连接移动设备（这种设备提供到处移动中的连接能力）到一个固定网络（此时用户处在家中）的关注力。这是服务重新分发的一个例子，可由家庭网络中的IP融合做到。使用IP，支持不同特点的设备间的融合。

在家庭网络之上数据重新分发价值的另一个例子可能是，将社交网络的访问扩展到位于起居室中的TV的可能性。在家庭中联网设备的数量将在接下来的数年间增加，将它们连接到一个共用网络的能力将催生新应用的机会。

在满足前面各例中描述的通信需求的目标下，不同设备必须具有被连接到远端单元的能力。将一条连接部署到每台设备的外部，将是一项巨大错误。最一致的战略是与外部共享单个连接点，在其上考虑安全性和可靠性问题，并由一个家庭网络（所谓的家庭骨干）互联在家庭中存在的所有设备。负责提供与外部世界连接的设备，被定义为家庭网关（或驻地网关），并将支持一种优化的尺度定制和到达/外发通信资源的管理。

5.2.3 IP 家庭中的服务重新分发

上面已经简短地描述了家庭网络演进中的强烈趋势。对向一个更泛在的、可访问的、灵活的和鲁棒的基础设施演进的需求，看来是清晰的。但在真实事实方面，一个全 IP 网络的主要方面，将是提供一个融合网络，它将支持在其上提供极多种类的服务。

1. 话音和电话服务

使用光和/无线电电子信号以一种双向方式将话音传递到远端位置，被称作电话。传统上而言，电话服务是通过 PSTN 基础设施由电信运营商提供的。

自数字化电话服务的首次发行以来，话音应用的涉及范围（谱系）已经朝 IP 上的话音（VoIP）概念演进。如可从名字推断出来的，VoIP 通信的主要优势是从所用的基础设施抽象出话音服务，提供一种物理媒介和协议融合，这要感谢 IP 网络层。VoIP 服务将在家庭骨干的异构网络分段之上分发，而与外部世界的连接将由 IP 协议驱动执行。

要被重新分发的流将需要大约数十千比特/秒的吞吐量，对所用编解码具有直接的依赖关系（G. 721、G. 723、G. 729、μ 律、a 律等）。

由于实时要求的行为和通信双方间的交互，话音重新分发服务对时延和抖动是敏感的。所以，底层基础设施必须以所要求的 QoS 响应在整个家庭上支持服务重新分发。

2. 因特网接入和物联网

因特网接入服务反映了这样的机会，即基于为家庭提供与任何远端位置的一条通信路径。将位于家庭的嵌入式系统集成到物联网，将必然改变未来家庭图景。数据事务（交易）、网页浏览和电子学习或电子健康服务将是新应用领域中的一些应用。

因特网服务要求越来越高的具有突发控制的传输速度。存在中等程度的时延是可接受的，错误和丢失的分组由不同 OSI 层中的协议加以处理。

对等（P2P）服务的日渐增多使用和视听内容流化极大地改变网络行为。在一个 P2P 网络中，不同节点可作为客户端或服务器甚至同时作为两种角色。没有可预测的固定节点行为。这种联网概念本质是随如下事实而逐渐浮现的，即端用户正成为内容生成方和消费者，在所连接的节点间有直接的内容交换。这将导致网络共享资源使用的强化新 QoS 范型的出现。

因特网服务的到达（和外发）流应该以对用户透明的方式在家庭骨干之上重新分发。出于这个目的，家庭网络集成流量优先级分级机制，管理有限的联网资源，将这些资源与日渐增多的服务数量进行共享（流化、万维网、电子邮件、VoIP 等）。

3. 电视服务（HDTV 蓬勃发展）

电视服务包括视听内容的传输，主要将焦点放在娱乐应用上。数字化电视重新分发，要求不同的带宽，这取决于所用的分辨率和在传输中所用的编码算法。因此，对于要被重新分发的每个节目（音频+视频+关联的数据）所使用的吞吐量，可在2~20Mbit/s之间变化。如今，在家庭中存在几台电视机是常见的，在放置于家庭不同位置的每个屏幕上具有不同的节目呈现。所以，通过家庭骨干的内容重新分发是需要的，无论在家庭还是扩展的家庭中都需要。

视听内容重新分发是家庭骨干中具有最高资源需求的服务之一。由此，由于服务本身的特征，将不同的优先级和QoS机制应用到不可中断的实时服务。事实上，视听内容重新分发一直是QoS研究领域的主要关键。前面提到的所需联网资源持续地随HD电视（HDTV）和三维电视（3D-TV）服务的流行而增加。

4. 交互式视频和多媒体内容流化

交互式视频服务使用户能够实时地发送和接收视听内容。用户端对这种服务的兴趣日渐增加，用户成为视听内容和服务的提供商。

第一代多媒体流化内容服务之一是视频会议服务，也称作IP上的视频和话音（V2IP）。与经典的数据传输相比，这项服务要求较高（带宽）的流。与视听内容共享服务（即YouTube、CNN iReport等）的巅峰流行相结合，其中突出的是，在将视听内容与我们的同事、邻居或位于遥远城市的朋友共享方面，存在日渐增加的关注度。

5. 家庭自动化服务

家庭自动化服务一直是家庭网络部署的首要正当理由之一。家庭骨干将在未来家庭的几项任务的自动化中扮演一个关键角色，如提高能源效率［绿色信息技术（IT）］的家庭仪器设备的自动化。

在过去数年间，每家家庭仪器设备制造商采用一种专用的联网技术，使用电力线或一种无线物理媒介（蓝牙、ZigBee等）作为物理基础设施。

在未来家庭中，每个设备都必须被连接到家庭骨干，支持不同系统间的互操作。

从带宽需求分析角度看，自动化服务过去要求低带宽通信。但是，由于未来家庭自动化命令的至关重要特征，必须考虑到它们与其他通信（在家庭骨干之上重新分发）的共存。

6. 环境辅助起居（AAL）服务

在未来将引入一代新服务，即所谓的环境智能（AmI）。这是在过去数年间由研究共同体定义的一个概念，但直到最近，可由绝大多数用户支付得起的第一批实用实现才出现。

但是，每项预测都预言，一种新的技术浪潮将我们的社会引入一个新时代，这是基于AmI服务的一个时代。以此为主线，业界、大学和研究中心正不断地引导

技术革命，并以所需的通信基础设施提供给市场，以便提供创新型的应用[1]。

家庭骨干基础设施将必须向用户提供一种泛在和弥散（无处不在）的环境，无论他或她位于家庭中还是扩展的家庭中，服务到达他或她的方式对用户是透明的。

5.2.4 全 IP 家庭骨干的容量

如在上面所述，依据要被集成到家庭骨干的不同服务之通信需求，将反应非常不同的需要（见图 5.2）。但是，在存在于家庭的每个联网设备间提供一个统一的连接点，看来是必要的。因此，在家庭联网演进中主要执行轴线之一，必须基于一种网络的改善，这种网络融合不同的不相交的网络分段（细分市场）。

前述融合的催化剂是，采用不同物理媒介和协议连接的各设备间的互操作能力，以这项需求为主线，IP 超越了其他各协议，将家庭骨干变换为一个全 IP 基础设施。



图 5.2　家庭骨干中不同服务的集成

在前面描述的服务考虑到与多媒体服务、数据网络、接入控制服务、家庭自动化和无线传感器网络有关的流，目标是感知家庭环境等。这些服务建立不同的需求类型：与家庭内服务重新分发有关的需求类型，以及与进出家庭的到达/外发流有关的需求类型。

表 5.1 给出了由家庭骨干提供的未来需求的一项估计（从吞吐量角度看）。

表 5.1　家庭骨干中服务所要求的吞吐量

服　　务	带宽/(Mbit/s)	设备数量	要重新分发的流/(Mbit/s)	到达带宽/(Mbit/s)	外发带宽/(Mbit/s)
TV 流化	2～25	3	75	50	0
数字录音机	2～25	1	25	0	25

（续）

服　　务	带宽/(Mbit/s)	设备数量	要重新分发的流/(Mbit/s)	到达带宽/(Mbit/s)	外发带宽/(Mbit/s)
家庭影院	1~25	1	25	0	0
因特网浏览器	1~20	5	20	20	6
视频会议	1~4	1	4	4	4
数字电话	0.2	5	1	1	1
网络游戏	0.2~2	3	6	6	6
视频监控	0.1~1	10	10	0	1
家庭自动化	1	8	8	0.2	0.2
便携音频	0.1~2	3	6	0	0
合计	—	—	~165	~85	~1~40

5.3 家庭（全IP）骨干

在前面一节中描述了由家庭骨干支持的不同服务的通信需求。下面将深入讨论家庭骨干架构的定义，目标是识别其关键特征。

在新应用服务领域的产生和新附加值服务的产生方面，服务的互操作能力和在连接到家庭骨干的不同设备间通信的能力，将扮演一个关键角色。

5.3.1 IP作为家庭骨干网络的关键实体

新的家庭联网框架突出了“物联网”的重要性，其中涉及互联位于家庭中大量联网的嵌入式系统。

在家庭网络中存在的各设备将要求宽范围的连接能力。将存在具有不同特征的不同节点，但必须建立一个共同的黏合点，以便服务所要求的互操作能力。为完成这项概念性融合，可使用OSI参考模型作为一项指南。



图5.3 全IP融合中的OSI参考模型

如图5.3所示，从通信观点看，OSI模型将一个节点的架构分为7层栈。依据所使用的通信技术，每个不同的网络分段具有低层的自己特定的实现。所以，在这两个低层中，每个异构网络分段有

所区别。

但是，通过在网络层提供一种同构的机制，提供 IP 融合的概念得以延续。这种机制将负责通过异构分段连接的嵌入式系统的互操作能力。

因此将提供所述融合的各单元，组成每种联网技术的特定物理层和链路层，并在 IP 层（在网络层）提供集成。采用这种方式得到一种全 IP 架构。

5.3.2 家庭网络相关的联网技术

接下来，将描述最重要联网技术的一个简短历史。重要的是强调，本章并不放弃任何联网技术，而是基于由不同网络分段的并集所组成的融合之上，提出补充性的愿景。

1. 以太网（IEEE 802.3）

可能的情况是，这项通信技术最可能扩展到整个世界。最典型的以太网速度是 100Mbit/s 和 1Gbit/s。它由两个低层 OSI 层栈［物理层，媒介访问层（MAC）+ 数据链路层］所表征。通常的情况是，IP 用在一个以太网栈之上。在一种特定 MAC 基础上，以太网成为最具扩展能力的数据网络[34]。尽管如此，标准以太网不是提供高 QoS 和硬实时服务的最适合联网技术。

大量研究精力已经将焦点放在适配以太网方面，使之用作一项多媒体内容重新分发技术。但是，我们将看到，关键点可能是与其他联网技术的融合，而这些技术是为这个目的而设计的。

2. 工业以太网（实时以太网）

依据实时服务分发的日渐增长需求和识别新的关键型的和确定性通信基础设施需要，出现工业以太网。基于传统以太网（IEEE 802.3），它以极低时延和超低抖动[32]，提出改进 QoS 响应的不同替代方法。

主要目标是在嵌入式设备间实时、安全关键型通信的确定性。采用这种方式，正常数据和关键型数据可共享相同通信网络，取得所要求的时序响应。一般而言，基于 IEEE 1588 协议上的等时性（Isochrony）机制。

不同方案（20 种以上）被考虑作为工业以太网：EtherCAT、EtherNet/IP、Profiner-IRT、Powerlink、TTEthernet 等。

3. IEEE 1394

也被称作 Firewire 和 i-Link，其等时性特征成为 IEEE 1394，作为分发 HD 服务的最合适的联网技术。它具有提供所需 QoS 的固有机制，为等时性通信保留高达 80% 的带宽，为异步通信保留 20%。IEEE 1394 在不同物理媒介之上提供高带宽解决方案（高达 3.2Gbit/s）：铜质走线、同轴、5 类、光纤等。

IEEE 1394 的即插即用特征为端用户提供了容易采用的便利。IEEE 1394 提供一种拷贝保护机制（IEEE 1394 DTCP），并被考虑作为车载娱乐的一种多媒体总线（IEEE 1394 IDB）。1394 TA（贸易联盟）[35] 处理无线扩展，称作无线-1394。

4. 电力线

电力走线被认为是在家庭中具有最高毛细血管渗透能力的走线。同时强调的是，电力线通信（PLC）不需要任何新的走线，原因是已经部署的走线被用作通信基础设施。无论用户希望在哪里安装一台新的设备（将被连接到电力供应），都将隐含着在电力线网络上200Mbit/s以上的一个通信接入点。

一条电力线的主要问题是媒介的不利特征[33]。它基于为电力传输而设计的一种物理媒介，所以它不被看作一种数据通信网络。

但是，感谢来自不同公司的支持（DS2、Intellon、Gigle、松下、Echelon、SpeedCom等），电力线已经成为用于家庭骨干的最有前景技术之一。在过去数年间出现了不同标准［统一电力线联盟（UPA）、Homeplug、HD-PLC］，最近美国电气电子工程师学会（IEEE）已经批准其IEEE P1901作为电力线网络上的宽带规范[36]。国际电信联盟（ITU）也将在ITU G. hn[37]考虑电力线媒介。

5. HomePNA

HomePNA使用的物理媒介是电话线。其哲学理念是使用现有走线（电话线），目的是在其上提供改进的带宽通信[38]。使用所述走线的主要问题是，这种走线在高频时遇到巨大衰减，所以使用自适应正交幅度调制（AQAM）。

在HomePNA中存在几次演进（HomePNA1. 0、HomePNA2. 0、HomePNA3. 0和HomePNA3. 1），速度高达320Mbit/s。最后提到的选项支持耦合到同轴线缆。

6. USB

USB为一种初始的外设连接（高达12Mbit/s）。USB2. 0达到480Mbit/s，如今USB3. 0设备已经在开发之中[39]。

一个USB支持热插拔，并被集成在多数电器设备［电话、个人计算机（PC）、TV、PVR等］内。与IEEE 1394的对比差异是，一个USB要求存在一个主机，它将管理该总线。这个问题可采用即插即用（USB-on-the-go）（OTG）解决方案加以解决。

在USB2. 0中考虑了同时性传递，但没有资源预留机制。因此，USB2. 0在QoS提供方面没有IEEE 1394那样好。

7. UWB和IEEE 802. 15. 3a

一个超宽带（UWB）系统被定义为这样一个通信系统，它由如下条件的任何一个条件所刻画：

1）在10 dB的带宽大于500MHz。

2）在10 dB处的带宽及其中心频率之间的比值大于20%。

自20世纪60年代，就知道了UWB技术。但是，其效用主要受限于军事应用，如脉冲雷达。如今，感谢电子器件价格的降低，UWB技术被看作是一种可支付得起的连接解决方案，用来提供无线短距离连接能力。

因为有数百Mbit/s的带宽，所以UWB被看作消费电子市场中关键技术之一，

它将HD机顶盒（STB）或硬盘驱动（HDD）连接到新的屏幕[2]。

个域网（PAN）和体域网（BAN）场景是UWB技术最有前景场景中的另外两个场景。

8. 无线USB

在PC产业中有导线USB解决方案的巨大成功，刺激工业界基于USB的无线扩展[40]开发一个标准化过程。

使用UWB技术。有几家通信设备制造商正在开发第一代无线USB设备，该设备提供480Mbit/s USB有线解决方案的一种无线扩展。HD内容流化也将是无线USB技术的关键市场之一。

9. 蓝牙

这是一项短距离和低功耗通信技术。通过将这项技术植入到PC、个人数字助理（PDA）、蜂窝电话、鼠标、键盘、打印机和免提设备等，这项技术的成功得到提升。

蓝牙得到蓝牙特殊兴趣组（SIG）的支持，并在1998年得到爱立信、英特尔、IBM、诺基亚和松下的大力提倡。它们开发了无线个域联网设备[41]的一种开放标准。

蓝牙组成一种同时型总线，但视频信道不能在蓝牙上重新分发，这是由于其低带宽造成的。但是，WiMedia联盟[2]正在基于UWB技术开发新的蓝牙3.0扩展。

10. IEEE 802.11a/b/g

也被称作Wi-Fi，IEEE 802.11a/b/g是比较广泛使用的无线网络技术家族，有很多制造商开发带有IEEE 802.11接口的产品（蜂窝电话、便签PC、膝上机等）。

IEEE 802.11提供数百米的通信距离，且这项技术的市场采用率是史无前例的[42]。Wi-Fi的设计是提供数据流连接能力，所以最初产生时没有传输多媒体流的足够QoS机制。尽管如此，研究共同体开发了几种方案提供QoS机制，因此在VoIP服务蓬勃发展内使用Wi-Fi作为一项关键技术。

IEEE 802.11提供的带宽达到数十兆比特/秒，但这对在家庭中几个HDTV频道的重新分发是不够的。

11. IEEE 802.11n

由于上述Wi-Fi的限制，Wi-Fi联盟产生了IEEE 802.11的一个同步分支，称作IEEE 802.11n[42]。在这项联网技术的设计中，在IEEE 802.11a/b/g网络上研究受到的限制在这方面有所帮助。

在IEEE 802.11n中使用的关键问题之一是，使用具有重要带宽增益的多输入多输出（MIMO）机制，这是由于使用分集技术得到的，达到数百兆比特/秒带宽。

12. IEEE 802.11e

IEEE 802.11e技术背后的工作组打算提供IEEE 802.11 MAC（目的是增加它的QoS）和物理层安全（从IEEE 802.11网络继承的）之上的一种新规范。主要问

题是，以一种面向时分多址（TDMA）的机制，替换MAC子层，并针对带优先级的流量，添加一个额外的错误纠正系统。

IEEE 802.11e背后的工作组考虑了IEEE 802.1p规范[42]，目的是提供分类流量流的一种方法，并为家庭内多媒体重新分发建立一个关键点。

13. ZigBee和IEEE 802.15.4

ZigBee是无线传感器网络场景中最广泛部署的技术。它有低功耗模式，被设计用作极低计算资源节点的通信接口[43]。

所选频带位于工业、科学和医疗（ISM）频带内，并在宽广范围的应用中替代蓝牙技术。ZigBee提供低数据速率通信，带宽可达250kbit/s。

ZigBee网络是在家庭联网基础设施中，集成无线传感器和家庭自动化网络的一个非常令人感兴趣的解决方案。在这项集成中，全IP融合层将扮演一个关键角色。

14. MoCA

同轴多媒体（同轴之上的多媒体，MoCA）联盟[3]是由几个工业界参与者（如Pulse-Link和NXP等）创建的，目的是使用同轴线缆提供家庭骨干基础设施。在正交频分复用（OFDM）基础上，MoCA使用50MHz传输信道传输高达270/400Mbit/s的数据流。在排队优先级和数据加密标准（DES）基础上，提供了QoS机制。

作为一个简短历史，突出强调的是，MoCA是在提供高数据速率家庭骨干方面，工业界努力工作付出的结果。

5.3.3 联网技术总结

在读完前面的内容之后，读者应该注意到，存在多种联网技术，在家庭骨干中有直接应用。但是，重要的是强调，没有预见到一种绝对占主导地位的通信技术。为了重新分发一项特定的服务，每项替代方案在成本、距离和QoS方面都进行了优化。

所以，考虑服务的异构性和对重新分发做出响应的需要，融合骨干再次成为解决方案的焦点。不同技术的并集，还有它们的通信资源，将是最吻合一致的家庭骨干。

在表5.2中，可看到一些联网技术特点的简介。

表5.2　家庭联网技术小结

	物理媒介	QoS（同步的）	距离范围	数据速率范围
以太网	5类、6类	否	百米	100Mbit/s，1Gbit/s
工业以太网	5类、6类	是	百米	100Mbit/s，1Gbit/s
IEEE 1394	光纤、同轴、铜线	是	百米	<3.2Gbit/s
光纤	光纤	是	千米	>10Gbit/s

（续）

	物理媒介	QoS（同步的）	距离范围	数据速率范围
USB	USB线缆	—	米	>480Mbit/s
IEEE 802.11g	无线	是	数十米	<54Mbit/s
IEEE 802.11n	无线	是	数米	300Mbit/s
UWB	无线	是	百米	480Mbit/s
ZigBee	无线	—	数十米	250kbit/s
蓝牙	无线	是	数十米	3~24Mbit/s
PLC	电力线	是	百米	200~500Mbit/s
HPNA	电话线	否	百米	320Mbit/s
MoCA	同轴	是	百米	400Mbit/s

5.4 家庭网关

虽然家庭网关的研究超出本章的范围，但给出（至少）家庭骨干入口/输出口的一个简短引用是基本要求。

家庭网关，也称作驻地网关，是负责完成家庭骨干和外部网络之间连接功能的设备。家庭网关提供骨干的一个抽象层，从而就服务到达家庭的方式方面，对设备是透明的[4,28,30]。

在家庭网关最有价值的功能中，可重点强调为家庭提供进/出连接能力、与不同网络分段的融合和为管理异构网络分段互操作能力而提供的QoS机制（见图5.4）。

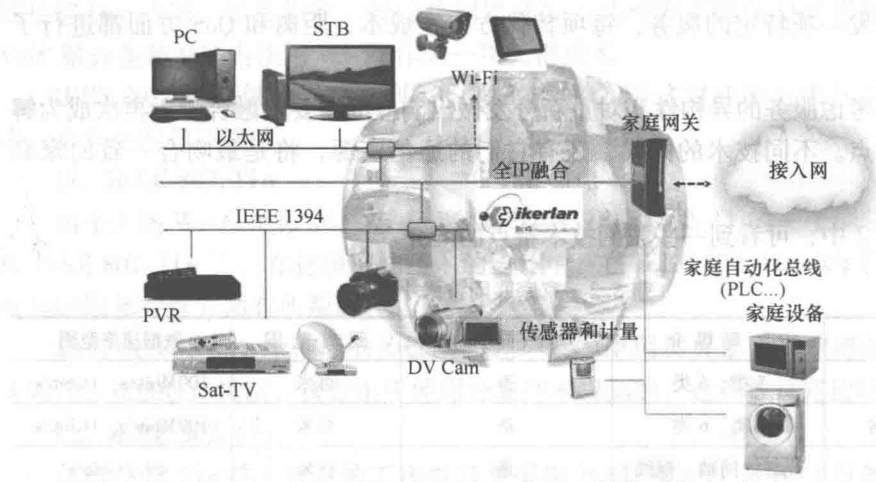


图5.4 作为外部网络和内部网络之间连接的家庭网关

由家庭网关实施的目标有接口灵活性、扩展性、安全性、可靠性和远程管理能力，目的是控制家庭骨干的资源。

围绕家庭网关单元的研究，可给出大量参考文献。但 Zaharadis 教授的书籍可作为一个不错的参考起点。

5.5　桥接各项技术：步向全IP基础设施

5.5.1　桥接全IP融合架构

三重播放和四重播放多媒体设备的融合需要，来自于当前出现了许多消费电器（如数字相机、嵌入式PC、蓝光设备、手持设备、HD机顶盒等），这些设备带有日渐增多的联网接口。

由家庭网络提供的附加值，准确地说是到达多媒体流和内容重新分发到远端显示和存储系统的混合体，这使它对满足未来数年数字家庭的用户需求是有吸引力的。多媒体技术方面（数量）的增长及提升的带宽需求，是由HD服务的引入带来的。基于在消费电器领域中不同的领先公司的市场分析，本章作者所实施的一项研究，突出了为将内容重新分发到家庭各处而需要有数百兆比特/秒的一个家庭骨干网。所以，为用户提供一个泛在多媒体环境所需的基础环境，不是那么简单的事情，这样的环境需要满足所需的容量和QoS。

对在家庭网络上传输的流量类型，必须要做一些考虑。一方面，异步数据传递将是常见情况，如文件下载、照片传递等，另一方面，将存在与视听内容特征有关的同步传递。多媒体传输通常由缓冲的使用和非实时可视化所表征，但必须考虑时延和抖动，以便在视听交互服务中满足所需的QoS。

研究共同体完成了一项认真的研究，目的是选择最优的总线，以便在家庭环境中互联多媒体设备。但是，工业界已经为用户家庭中存在的电器设备开发了不同的联网技术。这个事实说明，在家庭中存在不同的联网技术，但构成不相交的通信孤岛。每条通信流有自己的（通常是不同的）QoS需求，所以对每种场景，联网资源是不同的。异构性的另外一个原因是基础设施的成本。

在IP融合时代，出现了布置融合家庭骨干的机会，这个骨干由不同通信技术组成，这些技术有其相应的QoS响应，并提供各种机制（桥接），以便在相同的互联家庭骨干下连接每个异构分段。事实上，家庭骨干的目标是提供互联各设备的一种方式，这些设备是连接到不同特征的网络分段的。

最广泛部署的通信网络是所谓的以太网（IEEE 802.3）。由于历史原因和成本原因，在巨量设备家族中都集成了以太网接口。这是在家庭中用来扩展因特网连接能力的最典型的技术，在办公室和家庭数据网络环境中这种接口的存在是海量的。要强调的是，以太网栈上面的网络层是著名的IP，而且这个事实在促使IP成为融

合未来的中心方面，具有重要的分量。

所以，如果以太网是最广泛使用的技术，为什么我们不朝着一个基于以太网的家庭骨干演进呢？答案可能有一个积极的方面，但存在与以太网相关的几项限制。以太网的诞生，目标是在工作站间提供连接能力（主要针对的是文件传输）。尽管如此，在过去数年间，通信需求发生了指数增长，而且在数年间做出非常良好响应的联网技术，不再是高带宽消耗多媒体流重新分发的最合适技术，就像HDTV及其高QoS行为的情形。

传输高吞吐量、低时延和抖动流等的需要，意味着需要新的联网技术。IEEE 1394，也称作相线或i-Link，是围绕多媒体流重新分发需求为主线而设计的一项技术的例子。IEEE 1394支持同步流的传输，并具有带宽预留的固有QoS机制，其在视听内容流下的行为是非常奇特的。

在家庭骨干研究中的起点必须是流的分析，这些流要在前面提到的基础设施之上重新分发。一方面，异步数据传递将是常见情况，如文件下载、照片传递、网页浏览等，另一方面，将存在日渐增加的同步流量流，它们被重新分发到家庭各处。一般而言，视听世界事实上是同步的（包括HDTV服务）。在同步服务和异步服务间将出现共存，每种服务都要在一种不同联网技术上被传输，这些技术要适配服务的需求和支付得起的成本，目的是保障所要求的行为。

本章建议一种基于IP的基础设施，作为家庭骨干鱼骨型（fishbone）结构，这源于其互操作能力。尽管如此，这个骨干将由不同异构分段组成，这与IEEE 1394的情形相同，目的是在没有抖动（分组交换网络中固有的）的情况下，重新分发实时视听内容。所以，融合的扩展家庭网络架构将基于IP网络OSI层，同时链路和物理层将由不同的异构分段（PLC、IEEE 802.3、IEEE 802.11b、IEEE 802.15.3、IEEE 802.11n、Firewire或IEEE 1394等）组成。下面将重点描述在以太网（IEEE 802.3）和IEEE 1394网络分段之间桥接的例子，其目的是取得一种IP层融合的效果。

IEEE 1394总线被设计为消费电子设备的一个接口。IEEE 1394将可用带宽细分为20%专用于异步事务，其他80%用于同步流的重新分发。这个事实使IEEE 1394能够提供卓越的QoS响应。IEEE 1394降低了抖动，抖动以前是非同步网络技术的主要问题，这与以太网的情形相同（几种实时以太网变种是例外情况）。

IEEE 1394得到消费电子工业的支持，因此具有高的集成水平，被集成到HDTV设备、笔记本电脑和视频录像机等。其即插即用特征，支持具有高速和可扩展性能的热可插拔设备的互联。IEEE 1394也许是在家庭中重新分发视听内容的最合适的通信技术（见表5.3）。

以太网的大范围现存市场和穿透力以及其他联网技术的存在，意味着需要在家庭骨干中将两个世界联结起来。围绕同步网络分段和异步网络分段的融合，存在大

表 5.3　IEEE 802.3 和 IEEE 1394 的对比

特　征	IEEE 802.3 网络	IEEE 1394 网络
同步视频	异步网络	同步网络
可靠性	尽力而为	可靠性好
不允许时延	有时延	无时延
为抖动设置界限	抖动，没有设置界限	设定界限的抖动
无分组重复	可能重复	不允许重复
广播分发	单播/组播流化	是为广播设计的
QoS	不是固有就有的	是（带宽预留）
带宽	100/1000Mbit/s	800Mbit/s（高达 3.2Gbit/s）
媒介访问	冲突	带宽预留

量研究工作[5]。但是，最令人感兴趣的融合建议之一由 IEEE 1394 和以太网网络分段的联合体组成。采取这种方式，同一骨干提供 IP 世界［SIP、实时协议（RTP）、实时控制协议（RTCP）、实时流化协议（RTSP）等］的服务、典型 IP 网络发现机制［和统一插拔（UPnP）的情形一样］，并支持多媒体设备（通过 IEEE 1394 和以太网分段连接）间的互联。所以，同一骨干提供以太网和 IEEE 1394 分段的优势，并提供 IP 网络的互操作能力。

图 5.5 给出了这两个世界之间联合体（由一个网桥提供）的全局架构。可以看出，在不同 IEEE 1394 物理媒介（铜线、光纤）间的融合，并与以太网形成联合体，这是由一个内嵌的网桥（或网关）实现的。这个网桥是提供融合骨干方面的关键单元，它提供 IEEE 1394 和以太网的优点，因此支持 HDTV、VoIP 和 IPTV 服务的集成，同时支持 PC 或其他 IP 设备间的数据流互联。

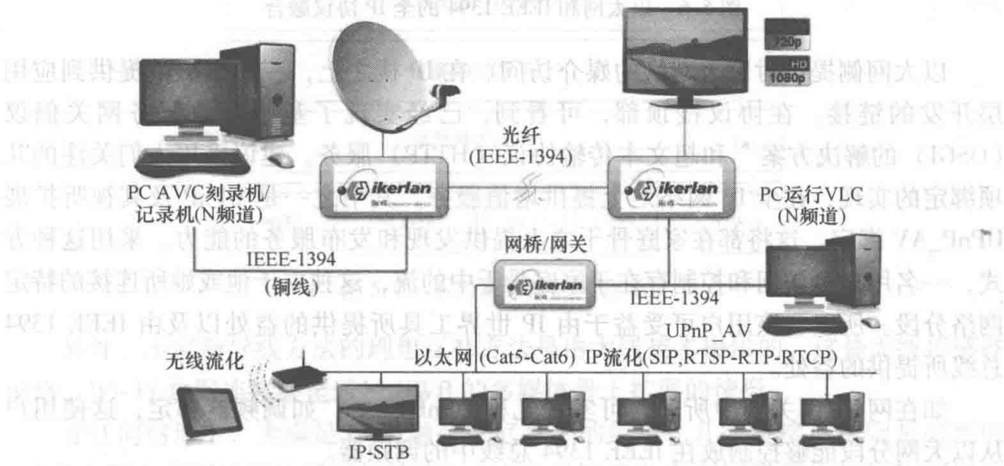


图 5.5　采用以太网和 IEEE 1394 分段的全 IP 家庭骨干融合

所述网桥由本章作者在参考文献［5］的工作中开发完成，目的是为用户的家庭骨干需求提供一种架构型的解决方案。构成网桥的内嵌系统，有一个 IEEE 1394 接口、一个以太网接口和协议栈（为向家庭骨干提供 IP 融合所需）。所以，连接到以太网分段的各设备，可访问由 HDTV 卫星调频器（通过 IEEE 1394 分段连接）接收到的内容。采用相同方式，连接到以太网分段的一台 IP-STB，可管理在 IEEE 1394 总线中接收到的流。另外，IEEE 1394 可基于 UPnP 访问 IP 网络中的服务发现机制，并可访问存在于以太网络中的流化内容。

桥接协议栈

负责在不同网络分段间提供融合（能力）的网桥，有双重功能。一方面，它必须支持物理层融合，以铜线、光纤、UTP-CAT6 布线等互联各分段。另一方面，相比于经典的网桥定义（网关功能），该网桥可完成增强的网关功能，这样就在 IP 网络层支持协议融合，因此建立了全 IP 场景。图 5.6 描述了要部署一个嵌入式系统（运行在 Linux 操作系统上）所需机制的实现，目的是提供协议融合。

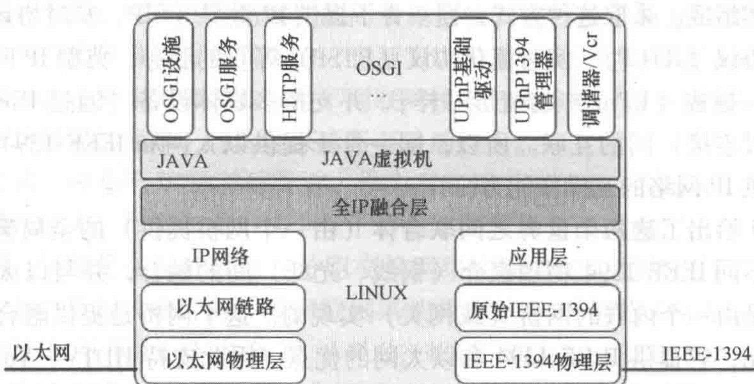


图 5.6　以太网和 IEEE 1394 的全 IP 协议融合

以太网侧提供对以太网络的媒介访问。在 IP 栈之上，一个虚拟机提供到应用层开发的链接。在协议栈顶部，可看到，已经实现了基于开放服务网关倡议（OSGI）的解决方案[6]和超文本传输协议（HTTP）服务。建议采用人们关注的几项绑定的实现，在全 IP 网络之上提供增值服务。其中之一是 UPnP 及其视听扩展 UPnP_AV 绑定。这将都在家庭骨干之上提供发现和发布服务的能力。采用这种方式，一名用户可访问和控制存在于家庭骨干中的流，这独立于他或她所连接的特定网络分段。所以，该用户可受益于由 IP 世界工具所提供的益处以及由 IEEE 1394 总线所提供的益处。

如在网桥/网关图中所述，可实现几种 UPnP 概要，如调频器绑定，这使用户从以太网分段能够控制放在 IEEE 1394 总线中的调频器。

通过 IEEE 1394 协议栈侧向下，可发现这样的协议层，是将 IP 分组净荷融合

汇聚到 IEEE 1394 中使用的同步总线和音频—视频协议（AV/C）。使用这个协议，控制同步侧的多媒体总线，因此在以太网和 IEEE 1394 网络分段之间融合汇聚控制命令。

5.5.2 不需新导线作为全 IP 基础设施的一种解决方案

在步向 AmI 的道路上，预计到以这样一种方式的基于内容和服务重新分发的一项杀手级应用，其中无论用户位于家庭的何处，他或她都将访问到该项服务。如已经描述过的，为布置一项合适的基础设施，需要部署与不同联网技术有关的布线。

围绕为向用户提供新的高增值服务的目标，为了建立新应用领域的基础，需要一项下层的家庭骨干基础设施。在需要一项基础设施（提供新服务）和新服务的存在（证明采用一项新的基础设施是合理的）之间，是一项循环论证，这是经典的鸡和蛋问题。

在所述环境中必须考虑一项附加的复杂度。在家庭中以及在公寓中，投入的全部责任和成本都落在端用户的肩上。

室外和室内世界之间的通信链路是由电信运营商提供的。但是，家庭骨干基础设施（支持服务和内容重新分发）必须由用户端买单，用户端同时是该项基础设施的拥有者。

这就是不需新导线（no-new-wires）范型诞生之处。新走线的成本是用户端采用该技术的一项重要障碍。出于这个原因，研究共同体已经实施开发通信技术的一项重要工作，这些技术将使用已经存在的基础设施，目标是利用由已经走线的基础设施［如电力线、同轴、电话线等的情形（见图 5.7）］所提供的可能性。

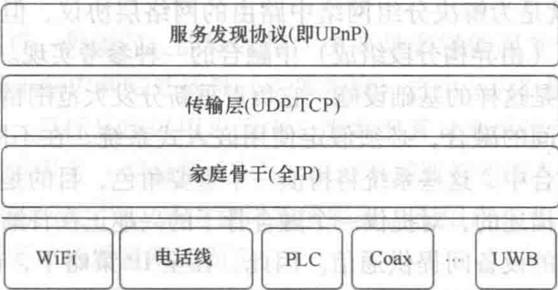


图 5.7　不需新导线的融合场景

另外，不需新导线方法的理想互补方法是由无线技术提供的。这是无线传感器网络、Wi-Fi 数据连接甚至通过 UWB 的多媒体骨干扩展的情形。

在任何情形中，为满足 IP 层融合家庭骨干的需求，几个桥接系统则是必要的（就像本章中前面描述的例子的情形）。这些系统将融合不同联网技术的物理层和协议。图 5.8 描述了在 IK4 联盟的一个雄心勃勃的项目框架中所开发架构的一个例

子，这个项目称作HOMI-IK4，翻译过来是智能多媒体家庭（IMHO-IK4）。可以看到，IP层提供一个以太网（IEEE 802.3）分段和UWB无线扩展分段之间所要求的融合。

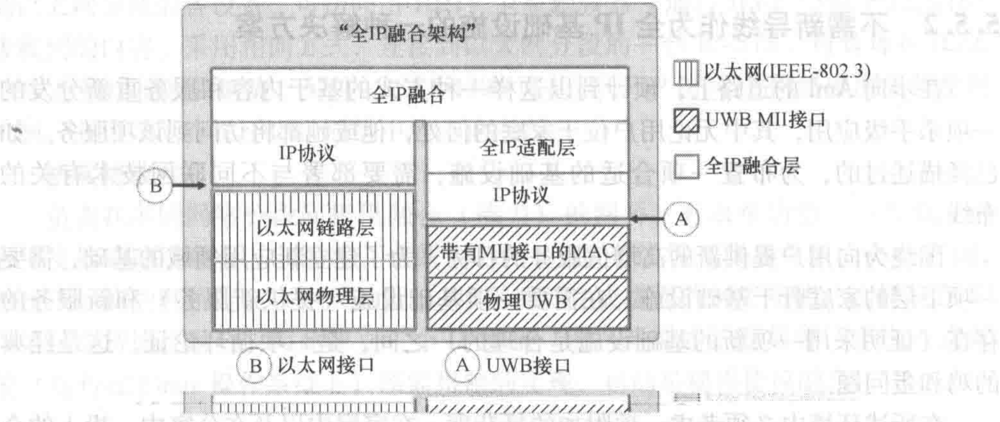


图5.8　以太网-UWB网桥的架构

所以，要感谢嵌入式系统［能够将信号耦合到现有导线（不干扰固有的功能）］和无线技术，在不需新导线准则下实现一个全IP融合家庭骨干才是可能的。

在参考文献［7］中，本章的作者提出一种新的QoS刻画方法论，来识别不需新导线的网络分段，这比较适合在家庭范围内重新分发一项给定服务，满足各项QoS需求。

5.5.3　物理媒介和协议融合

虽然IP生来就是为解决分组网络中路由的网络层协议，但它变成了一个OSI层，作为取得网络（由异构分段组成）中融合的一种参考实现。

家庭骨干必须是这样的基础设施，它负责重新分发大范围精彩纷呈的服务。为取得异构网络分段间的融合，必须假定使用嵌入式系统。在不同（网络）分段的物理联合和逻辑联合中，这些系统将扮演一个重要角色，目的是提供协议融合。

如在本章做过描述的，对提供一个融合骨干的兴趣正在日渐增加，目标是在带有不同特征之接口的设备间提供通信。因此，在全IP策略下，部署将在物理层和链路层建立融合的设备单元，就是必要的。

主要目标是完成一项通用基础设施，该设施将支持任何种类的服务，更具体而言是支持要求严格QoS行为的服务。协议融合基于如下挑战，即利用由不同协议提供的各种机制，目标是将每种机制的特征映射到其他机制上。

在图5.9中，可观察到一个语境图，它恢复由不同特征网络分段（针对每种情况，使用最合适的物理媒介）、高毛细现象、泛在性、自适应性、用户友好行为、可信赖性和可靠性的共存而混合构成的一个架构。

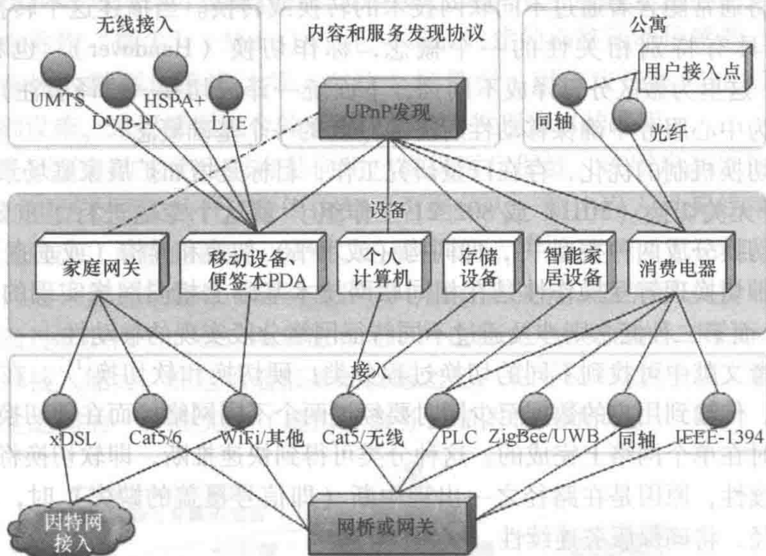


图 5.9　融合家庭骨干的参考架构

5.6　全 IP 家庭网络基础设施之上的服务

给出 IMHO 的两个不同场景，目的是将对扩展家庭骨干的关注点聚焦在以用户为中心的新服务提供和新商务模型的产生的框架上[30]。

5.6.1　扩展家庭之上的随身使用四重播放服务

如今，存在这样一种倾向，即将内容（主要是多媒体服务流）重新分发到移动中的用户[8]［处于室内和室外场景（扩展家庭）之中］。将通过有线网络或无线网络互联几种设备，目标是在以用户为中心服务部署日渐关注的框架[10]中，为到处移动中的用户提供服务（四重播放）[9]。用户不需要搜索服务访问点，但无论用户处于扩展家庭中的何处[11]，都将看到各项服务，在用户（他或她）可达的情况下，内容都将以所要求的质量跟随而至。

在定位机制方面的进展，为步向以用户为中心的新服务打开了一扇新的大门[12]。这些新的定位方法将帮助家庭骨干定位用户，并为用户提供内容和服务，从而他或她将选择他或她想在哪里消费内容和服务。在这种类型的场景中，至少需要如下角色：能够跟踪用户的一项服务，至少一种用户定位机制，一个 IP 融合家庭骨干和一个系统，该系统可告知内容服务器它必须将传输重定向到哪里。

在以用户为中心服务提供中的关键问题之一是这样的事实，即这些服务是在融合异构网络骨干上提供的，该骨干基于不同接入技术、协议和物理媒介。所以，用

户移动性将通常隐含着通过不同联网技术的转换或转换。当描述这个转换过程时，它看来是具有特别相关性的一个概念，称作切换（Handover）。也称作转递（Handoff，这里为做区分，译成不同词，下面统一译成切换——译者注），这是为在以用户为中心服务中确保移动性所需的 QoS 的一个基础概念[13]。

围绕切换机制的优化，存在巨量研究工作，目标是增加扩展家庭场景服务中的 QoS。媒介无关切换（MIH）或 802.21 工作组[14]就这个专题进行了重要的研究。工作组将切换分成两种类型[15]，即同构（或水平）切换和异构（或垂直）切换。

第一种切换理解互操作性是在相同联网技术基础上相同网络实现的（局部化移动性），而第二种概念则涉及通过不同特征网络分段实现的移动性。

在参考文献中可找到不同的切换过程分类：硬切换和软切换[13]。在一个软切换过程中，传输到用户的数据至少同时要经过两个不同网络，而在硬切换中，数据分发是同时在单个网络上完成的。这种分类可得到快速推断，即软切换将提供较高的服务连续性，原因是在路径之一出现中断（即信号覆盖的缺失）时，存在另一条下层路径，将确保服务连续性。

人们追求切换机制间的融合，目标是无论用户位于哪里，就像他们处在起居室中一样，向他们提供流化服务。这个事实必须对端用户是透明的，服务不应遇到任何中断。

切换是 IP 研究世界中的热点专题之一，存在巨量活跃的研究工作，建议使用 SIP 来提供随身服务移动性[16]。但是，基于 SIP 的切换机制遇到一项人们所不期望的时延，在切换过程中出现信息分组的丢失，这降低了 QoS。存在新研究的数据流，它们的关注点在这项 QoS 保障上。

一般而言，为以最合适的连接能力识别家庭骨干分段，切换过程是由信噪比（SNR）检测发起的。存在几种状况，即在一个拥塞状况中，切换过程是至关重要的。存在另一个研究动向，不仅基于 SNR 而且基于无缝切换过程的位置信息。在参考文献［17］中，Sulander 教授描述了 IPv6 移动网络上的另一种有趣的快速切换方法。

所以，存在不同的切换方法，但所有这些方法都共享这样的主要目标：无论用户端位于扩展家庭的何处，都向他或她提供不中断的四重播放服务。

在随身服务中，多媒体内容分发必须适配到用户的位置，要求不同的定位方法。这意味着，为定位用户，可使用不同的检测方法甚至混合方法，因此为用户提供多媒体服务。最常用的一些定位方法是全球定位系统（GPS）[18]、射频识别（RFID）[19]、UWB[20]、Wi-Fi[21]、ZigBee[22]等。

所以，用户定位方法以及异构网络上的垂直和软切换过程，使对扩展家庭 IP 融合骨干的关注变得突出明显了[29,31]。

5.6.2 e-健康应用

有特色的第二个场景，将焦点放在使用 IP 融合家庭骨干，在家庭中提供 e-健

康服务。在AmI时代，可想象这样一个场景，其中用户可购买几个小型嵌入式无线系统［也称作“微尘”（Motes）］，它可集成内建的心脏活动传感器、温度传感器、检测落体的微型加速度计等。在这个场景中，用户应该能够在家庭中部署e-健康基础设施，这要感谢融合的基础设施和自发现协议的使用。

这种类型服务的主要目标将是向用户提供医疗保健，该用户是单独（以一种泛在方式）生活在他或她自己房间内的一名老年人。前述嵌入式无线微尘将检测可能的紧急状况（可能的摔倒、晕厥、心脏病等），并将用户（无论该用户在扩展家庭中的何处）通过家庭骨干和驻地网关与医疗队连接，医疗队负责跟踪他或她的位置。通过家庭网关，医疗队和患者之间的这种交互是可能的，网关将为家庭骨干提供外部接入能力。在这项通信基础设施之上，采用HD视频会议，就可能激活患者和医生之间的一种准实时（虚拟）服务，如图5.10所示。



图5.10　电子健康应用场景

由一个融合的家庭骨干提供的主要优势之一是，它使用户能够获取（访问）传感器单元（ZigBee、RFID等），这些单元将度量和记录天气信息（即温度）、患者的状况等，并将这个信息提供给存在于家庭内的所有其他节点。一旦这些微尘连接到一个全IP融合骨干，则不同节点将能够自配置和检测所提供的服务。例如，位于起居室内的一台电视将发现一个微尘提供的服务，微尘提供所测量的温度（通过UPnP），从而可显示在电视屏幕上。感谢全IP融合部署，不同通信特征设备间的这种类型的交互通信才是可能的。

存在一个融合网络，应该会诞生基于AmI的新服务。采用一个泛在的、异构的和低成本的基础设施，用户比较容易采用新颖公众服务，这些服务如电子健康、电子行政、电子政府、电子娱乐、电子学习等，这有利于降低社会的“数字鸿沟”。IP融合网络将各项挑战转换为互操作性和新服务产生的机遇。

5.6.3　隐私、安全和用户概要

一个扩展家庭骨干意味着一个巨大的毛细网络，在用户端（无论他或她处于

扩展家庭中的何处）和服务提供商之间建立一条连接是可能的。但是，也出现了与连接型相关联的一个问题：在网络中可访问内容的安全和隐私问题。

一个融合网络向用户提供到服务提供商的端到端连接能力。因此，必须定义用户概要，从而一名用户仅可发现和访问这样的服务，其中他或她的概要被赋予访问权限。

骨干必须针对适宜（正确）的性能，部署可靠性机制，在没有所需机密信息条件下，避免不希望的由用户实施的内容访问，并保持内容隐私。所以，重要的是要考虑到与用户概要相关的安全和隐私问题。如果考虑一个 AmI 环境，则有必要定义对应于不同用户概要的多个访问身份，并协调处理隐私性和安全性[23]。

与全 IP 骨干相关联的中间件必须能够提供隐私性和安全性，从与中间件的机制相关联的复杂性，抽象用户端。所以，该用户必须能够以一种可信赖的方式，访问所需的服务（特别在关键数据相关的应用，如电子健康、电子政府等），对被重新分发的服务流，必须保障内容提供商的数据权益管理（DRM）。家庭骨干将以一种高效方式，管理不同的用户概要，前提条件是任何种类的数据流可重新分发到扩展家庭的各处，但仅有合格的用户才应该得到访问权限。

在共享的骨干中依据与隐私机制相关的每个问题，围绕这个专题，存在非常密集的研究活动。这些工作中的一些例子可见参考文献［24-26］。

5.7 扩展家庭网络

5.7.1 全 IP 融合网络上的垂直和水平传输

预计控制和多媒体流之间的融合是一项日渐急迫的需求。事实上，实时控制和多媒体流合并仍然是一项开放的挑战。从一个方面来说，人们期望使用相同的通信基础设施重新分发多媒体内容，并传输关键的控制命令。另一方面来说，不同网络分段间的互操作性和 QoS 问题是一个关键点，原因是这些分段将不再是不相交的。事实上，垂直和水平传输网络在扩展家庭全 IP 骨干中将扮演一个关键角色。

用于垂直运输（即电梯）和水平运输（即铁路）中的网络分段考虑采用创新的通信技术，称作工业以太网或实时以太网。在现场水平（焦点是控制问题）和多媒体设备中取得传感器和执行器间融合方面面临人们日渐增加的关注中，这些网络提供所要求的行为，其中多媒体设备面向交付信息娱乐服务，无论用户处于扩展家庭网络中的何处，均能到达他或她。

此外，在一个融合网络的体制中，全 IP 网络融合扮演一个关键角色，目的是为用户提供访问内容和服务的可能性，就像他或她位于他或她的家中一样。

为在相同总线之上取得控制数据流和多媒体的融合，人们提出了不同的解决方案，这些方案主要是基于实时以太网分段上的解决方案。在这些工业以太网之上，

取得IP层融合。实时工业总线建议间的主要差异在OSI低层（OSI层1和层2），但它们在IP层（OSI层3）是融合的，作为一个自然扩展，提供直接连接到IP家庭骨干的能力。

这些工业以太网技术的一些例子有EtherCAT、Powerlink（电力线）、Profinet-IRT、EtherNet/IP、Modbus-TCP等。

所以，在全IP框架下，通过使用联网技术，家庭骨干可得到扩展。图5.11给出了使用一个工业以太网络分段作为扩展家庭骨干组成部分的一个例子。

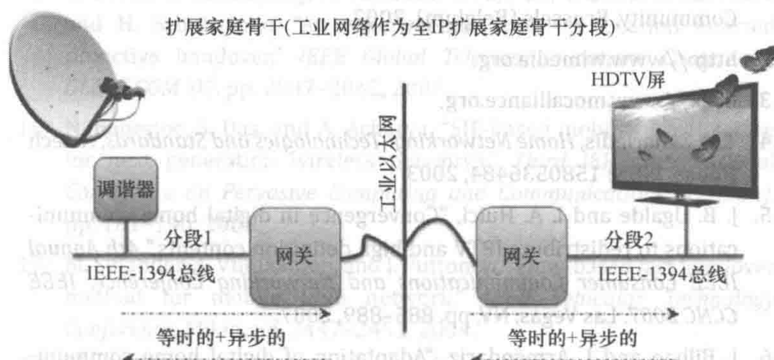


图5.11　工业以太网作为扩展家庭骨干的组成部分

5.7.2　全IP扩展家庭基础设施中的QoS

QoS是这样一个概念，针对一项特定服务，定义用来证明一种给定联网技术性能优良的合理性。QoS由一个度量指标组成，目标是衡量网络行为，并对一个通信网络的适用性进行分类，它是适用于任何扩展家庭网络分段的一个概念。

QoS量化背后的基础可概述为与传输有关的时延（延迟）、时延的变化（抖动）、吞吐量和比特错误率。在考虑所需吞吐量以及最大时延和抖动的情况下，所述QoS参数的测量和分析指明，一个给定的网络是否支持一种特定服务的重新分发。

从历史角度看，QoS分析一直与多媒体服务有关，这主要是由于视听服务的高需求（相比数据网络而言的）。但是，本章突出强调，新的扩展家庭骨干网络将必须遵循控制和多媒体分段的融合（准则），从而QoS保障到达一个重要的状态。

围绕QoS提供，存在巨量的研究工作。尽管如此，仍然可重点讨论由本章作者提出的QoS表征方法论[7]。所建议的方法论提供各种工具，分析一个给定网络分段是否是满足每种信息流所要求QoS的最合适网络分段。这种方法论帮助确定最合适的网络分段，用来在扩展家庭各处重新分发不同的信息流，满足所要求的QoS行为。

下一个10年，最重要挑战之一将是研究在共享资源网络上提供QoS的新方

式[45,46]，目的是能够确保所要求的吞吐量、时延、抖动和比特错误率，以使一个给定的流可重新分发到扩展家庭的各处[27,44]。

参考文献

1. J. Bilbao and I. Armendariz. "Consultation on the future of digital multimedia services and broadcast technologies," CORDIS, Definition of the VII Framework Program of the European Community, Brussels (Belgium), 2007.
2. http://www.wimedia.org.
3. http://www.mocalliance.org.
4. T. B. Zahariadis, *Home Networking: Technologies and Standards*, Artech House, ISBN: 1580536484, 2003.
5. J. B. Ugalde and I. A. Huici, "Convergence in digital home communications to redistribute IPTV and high definition contents," *4th Annual IEEE Consumer Communications and Networking Conference, IEEE CCNC 2007*. Las Vegas, NV, pp. 885–889, 2007.
6. J. Bilbao and I. Armendariz, "Adaptation of digital home communications infrastructure to carry IPTV and high definition content (HDTV) in the triple-play era," *45th FITCE Journal*, ISSN: 1106-2975, pp. 268–273, Athens, Greece, 2006.
7. J. Bilbao, et al. "Formulation and methodology for the analysis of viability of communication technologies in high QoS requirements multimedia flow redistribution (HDTV) in the extended-home environment," *IEEE BTS (International Symposium on Broadband Multimedia Systems and Broadcasting, BMSB)*, ISBN: 978-142442591-4, May 2009.
8. K. Kashibuchi, T. Taleb, A. Jamalipour, Y. Nemoto, and N. Kato, "A new smooth handoff scheme for mobile multimedia streaming using RTP dummy packets and RTCP explicit handoff notification," *IEEE Wireless Communications and Networking Conference*, WCNC, Las Vegas, NV, pp. 2162–2167, 2006.
9. H. Park, I. Lee, T. Hwang, and N. Kim, "Architecture of home gateway for device collaboration in extended home space," *IEEE Transactions on Consumer Electronics*, **54(4)**, pp. 1692–1697, November 2008.
10. H. Izumikawa, T. Fukuhara, T. Matsunaka, and K. A. S. K. Sugiyama, "User-centric seamless handover scheme for real-time applications," Personal, indoor and mobile radio communications, PIMRC 2007, IEEE 18th International Symposium, pp. 1–5, 2007.
11. S. Wang, S. Sridhar, and M. Green, "Adaptive soft handoff method using mobile location information," *IEEE 55th Vehicular Technology Conference*, VTC, **4**, pp. 1936–1940, Spring 2002.

12. K. Muthukrishnan, M. Lijding, and P. Havinga, *Towards Smart Surroundings: Enabling Techniques and Technologies for Localization*, pp. 350–362, Heidelberg, 2005.
13. G. Cunningham, S. Murphy, L. Murphy, and P. Perry, "Seamless handover of streamed video over UDP between wireless LANs," *Second IEEE Consumer Communications and Networking Conference (CCNC)*, pp. 284–289, 2005.
14. IEEE 802.21, http://ieee802.org/21/.
15. A. Dutta, S. Chakravarty, K. Taniuchi, V. Fajardo, Y. Ohba, D. Famolari, and H. Schulzrinne, "An experimental study of location assisted proactive handover," *IEEE Global Telecommunications Conference, GLOBECOM '07*, pp. 2037–2042, 2007.
16. N. Banerjee, S. Das, and A. Acharya, "SIP-based mobility architecture for next generation wireless networks," *Third IEEE International Conference on Pervasive Computing and Communications (PerCom)*, pp. 181–190, 2005.
17. M. Sulander, A. Viinikainen, and J. Puttonen, "Flow-based fast handover method for mobile IPv6 network," *IEEE Vehicular Technology Conference*, Milan, pp. 2447–2451, 2004.
18. "Standard positioning service specification," https://gps.afspc.af.mil/gpsoc/Default.aspx.
19. S. Manapure, H. Darabi, V. Patel, and P. Banerjee, "A comparative study of radio frequency-based indoor location sensing systems," *Proceedings of IEEE ICNSC*, pp. 1265–1270, March 2004.
20. S. Roy, J. Foerster, V. Somayazulu, and D. Leeper, "Ultrawideband radio design: the promise of high-speed, short range wireless connectivity," *Proceedings of the IEEE*, **92(2)**, pp. 295–311, February 2004.
21. "Wi-Fi technology," http://www.wi-fi.org
22. C. H. Lin, K. T. Song, S. P. Kuo, Y. C. Tseng, and Y. J. Kuo, "Visualization design for location-aware services," *Conference proceedings—IEEE International Conference on Systems, Man and Cybernetics*, Taipei, pp. 4380–4385, 2007.
23. L. Beslay and Y. Punie, "The virtual residence: identity, privacy and security," *The IPTS Report*, special issue on identity and privacy, September 17–23, 2002.
24. University of Taiwan: Ubicom Laboratory, http://mll.csie.ntu.edu.tw/index.php,2007.
25. Department of Computer Science and Engineering, Arizona State University, http://dpse.asu.edu/rcsm, 2008.
26. Pervasive Computing Laboratory, Universidad Carlos III, http://karajan.it.uc3m.es:9673/pervasive, Madrid, 2008.
27. J. Bilbao, I. Armendariz, and P. Crespo, "Disruptive mechanism in

the QoS provision," 49th FITCE International Congress, Santiago de Compostela, Spain, September 2010.

28. J. Bilbao, *et al.*, "Extended-home networks 2.0: guest starring actor in the social transformation facing the economical crisis," 48th FITCE Congress and *FITCE Journal*, Prague (Czech Republic), 2009.
29. J. Parra, J. Bilbao, A. Urbieta, and E. Azketa, "Standard multimedia protocols for localization in seamless handover applications," *Advances in Soft Computing*, **51**, pp. 191–200, Springer Berlin/ Heidelberg, 2009.
30. J. Bilbao and I. Armendariz, "Transforming homes: towards a heterogeneous user centric scenario, new opportunities and challenges," FITCE 47th International Congress, pp. 93–98, London, 2008.
31. E. Azketa, J. Parra, J. Bilbao, and A. Urbieta, "Standard multimedia for localization in seamless handover applications," 3rd Symposium of Ubiquitous Computing and Ambient Intelligence (UCAMI), Salamanca, Spain, 2008.
32. J. Bilbao, M. J. Zorrilla, G. Epelde, J. M. Perez, and I. Armendariz, "Industrial Ethernet architectures to provide QoS and add a convergent prospect for in-home high definition multimedia," *FITCE Journal*, 46th edition, ISSN: 1106-2975, Warsaw, Poland, August 2007.
33. I. Val, F. J. Casajus, J. Bilbao, and A. Arriola, "Measuring and modeling and indoor powerline channel," International Symposium on Performance Evaluation of Computer and Telecommunication Systems, IEEE Spects, San Diego, CA, 2007.
34. "Ethernet (IEEE 802.3)," IEEE 802.3 Ethernet Working Group, www.ieee802.org/3/.
35. "Trade association," IEEE 1394, www.1394ta.org.
36. "Standard for broadband over powerline networks: MAC and PHY layer specification," IEEE P1901, http://grouper.ieee.org/groups/1901/.
37. "Next generation home networking transceivers," ITU G.hn., http://www.itu.int/itu-t/aap/AAPRecDetails.aspx?AAPSeqNo=1853.
38. HomePNA Alliance, www.homepna.org.
39. USB, "Universal serial bus forum," www.usb.org.
40. "Wireless USB," www.usb.org/developers/wusb.
41. Bluetooth SIG, www.bluetooth.com.
42. "International standards for wireless local area networks," www.ieee802.org/11/
43. ZigBee Alliance, www.zigbee.org.
44. C. Cruces, J. Bilbao, and I. Armendariz, "Metodología para la caracterización de la Calidad de Servicio (QoS) de Redes Industriales

de altas exigencias," SAAEI-2010, Seminario Anual de Automática, Electrónica Industrial e Instrumentación, Bilbao (Spain), July 2010.

45. J. Bilbao, A. Calvo, I. Armendariz, and P. Crespo, "On High constraints over shared resource networks," *Proceedings—7th ACM/IEEE Symposium on Architectures for Networking and Communications Systems, ANCS*, pp. 221–222, 2011.

46. J. Bilbao, I. Armendariz, and P. Crespo, "Design of a new Queue management mechanism to optimize the use of shared resources in real-time interactive multimedia services," *IEEE International Symposium on Broadband Multimedia Systems and Broadcasting, BMSB*, Erlangen (Germany), June 2011.

第 6 章　无线车载网：架构、协议和标准

Rola Naja

6.1　引言

在不久的将来，因特网协议（IP）将从传统有线和无线网络的数据交换扩展到智能无线车载网络。事实上，信息技术和无线网络方面的进展，支持基于 IP 的多媒体服务和安保应用集成到车辆之中，使从车辆内的智能节点到因特网上的中央服务器的数据传递成为可能。

智能运输系统（ITS）目前是轿车制造商以及运输权威机构和通信组织关注的中心。ITS 被识别为这样的一项关键技术，可提升增进的安保、改善国家运输基础设施，并向道路用户提供至关重要的安保信息。因为多种类型的信息（即紧急消息、富媒体内容、信息娱乐数据等）在车辆和路侧集成设施之间交换，所以车辆到车辆（V2V）和车辆到基础设施（V2I）通信正成为一个 ITS 的两个重要组成部分。

无线车辆网络基础设施的部署，是正在进行的车辆到基础设施倡议（VII）主要焦点领域之一，这是在公众［联邦政府、州、本地政府、税务（toll）等］和私有（机车公司、ITS 设备制造商、通信公司等）干系方之间的一项联合行动。

车辆网络部署的主要目标可被总结为两个重要点：

1）最大化正面特征。确实，车辆联网将为司机提供本质的和有用的信息，用来增加人类和商品的移动性；改进驾驶舒适度。

2）最小化负面特征。更具体而言，车辆网络将通过应用防御性的碰撞避免技术，降低事故；通过将车辆流量限制在某个程度，降低拥堵；降低环境影响，这是影响来自于由各车辆实施的一项联合工作［计算碳排放］。

简言之，一个车辆间通信网络和一个 V2I 网络，在道路安全、检测和避免交通事故、降低交通堵塞以及改善驾驶舒适度方面，实施至关重要的功能。在此语境中，为了针对最小化车辆碰撞而提供基础工作，对无线车辆网络架构和服务质量（QoS）机制的深度理解，就是必要的。

因此，在本章将形成未来宽带车辆网络设计的一些深邃见解，这些网络能够适应变化的车辆交通状况和可变的移动性模式。更具体而言，将焦点放在车辆网络标准和车辆应用（在下一代车辆无线网络中预想到的）上。

本章组织如下。下一节重点讨论汽车安全领域中的主动和被动安全的概念，并描述目标为在一次碰撞之前和碰撞期间保护乘客的各组件。

6.3 节描述车辆网络架构，并给出在无线车辆网络内支持信息交换的主要设备。之后，说明车辆通信的不同类型和特征。

在 6.4 节，给出由车辆网络支持的多项应用和服务。基本上来说，描述两种类型的应用：安全应用和非安全应用，它们具有不同的需求和时延关键方面的特征。

6.5 节专门研究为车辆通信提出的各种标准。更具体而言，分析了陆地移动的通信接入（CALM）、车到车的通信联盟（C2C-CC）和车辆环境的无线接入（WAVE）。

在 6.6 节中重点讨论处理无线车辆网络研究工作的一些观点和挑战。最后，6.7 节给出本章的一个总结论。

6.2 实施主动安全

在汽车安全世界中，术语“主动”和“被动”是简单的但重要的术语。主动安全用来指在一次碰撞的防御中起辅助作用的技术，而被动安全指在一次碰撞期间帮助保护乘车人员的车辆的部件（主要有车辆的气囊、安全带和物理结构）。

由欧盟委员会为 2010 年设置的道路安全的运输政策目标，仅在采用一个集成的和整体性的方法（通过信息和通信技术）时才可达到。研究应该不仅将焦点放在碰撞阶段和碰撞后阶段，而且要将焦点放在碰撞前阶段上，要考虑被动、主动和预防性的安全措施。

在最小化碰撞风险中，预防性的和主动性的安全扮演一个重要角色。针对这一点，在车辆网络的语境中，人们提出了新的道路安全机制。

仅在定义安全功能、集成车辆内系统，并将它们与增强的远程信息技术组合到一个无线车辆网络中时，才可能做到安全系统的开发和快速传播。事实上，图 6.1 表明，在碰撞前阶段，车辆间通信的性能优于蜂窝通信。在碰撞后系统中，被动安全是由能量吸收措施、紧急呼叫、救援系统和服务提供的。而在碰撞前系统中，在将未来危险道路状况、容易犯错的司机、停止信号、紧急刹车、车道改变、前碰撞、交叉碰撞等信息提醒司机方面，车辆通信扮演一个重要角色。

关键点是，车辆网传播与救援、告警和信息应用有关的紧急消息。这些安全消息提醒司机，并对被救助司机所采取的行动具有至关重要的影响。

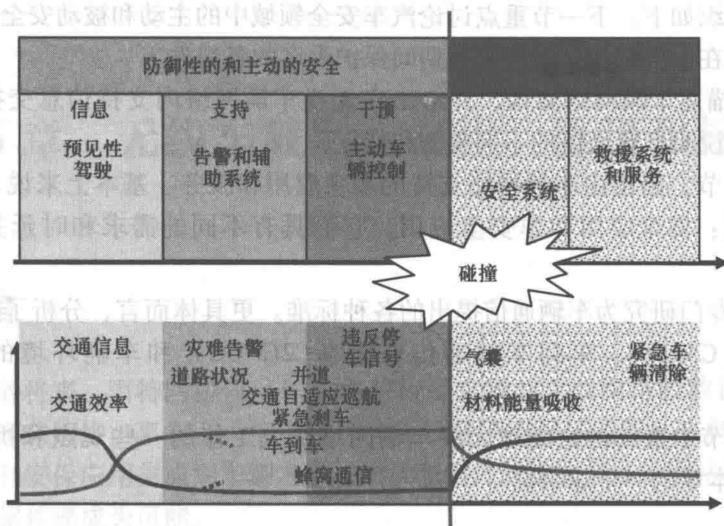


图 6.1　主动安全和被动安全（参见 Hitachi Europe，Sophia Antipolis）

6.3　车辆网络架构

6.3.1　智能车辆

ITS VII 尝试为“智能”车辆提供资金，方法是鼓励公私合作方，其中无线通信设备安装在国家的车队（私人投资），路侧通信基础设施沿运输系统的高速路、干线公路和十字路口安装。

计算和无线通信技术中的进展已经使人们增加了对智能车辆的关注：车辆装备有高级设备，这些设备为旅行者提供服务。智能车辆可被用来提高驾驶安全性和舒适度，并优化地面运输系统。

在最低限度下，一辆装备精良的车辆，配备有车载计算、无线通信设备和一台全球定位系统（GPS）设备，这使该车辆可跟踪它的空间和时间轨迹。车辆装备也可能包括一个预存储的数字地图和记录碰撞的传感器、发动机运行参数等。

在一个无线车辆网络中协作运行的未来车辆有许多特征：

1）车辆是内容的重要消费者：在每次旅行期间，乘客构成大量数据的被吸引的受众。例子包括位置感知的信息（基于地图的方向）和娱乐内容（流化电影、音乐和广告）。

2）车辆是内容的生产者：车辆报告道路状况和事故、交通拥塞监测和紧急邻居提醒（例如“我的刹车不工作了”）。

3）车辆是数据中继节点：事实上，所有应用都依赖于车辆（处在一个中介角

色）。在一个移动组设置中的个体车辆，协作改进整个网络的应用体验质量，为其他车辆提供临时的存储（缓存），并转发数据和对数据的查询。应该设计一种特殊类型的移动自组织网络（MANET）路由，支持车辆节点间传播消息。

6.3.2　路侧单元和车载单元

车辆架构支持两种类型的设备：路侧单元（RSU）和车载单元（OBU）。

1）一个 RSU 是车辆环境中的一个无线接入设备，仅当静态时，该设备才工作，它支持与 OBU 的信息交换。通常情况下，沿道路运输网络安装这样的设备。

2）一个 OBU 是一个移动或便携无线设备，支持与 RSU 和其他 OBU 的信息交换，并可在运动时工作。

6.3.3　车辆通信

车辆间的无线通信以及车辆和路侧基础设施之间的通信代表了一种重要类型的车辆通信。在车辆网络中，存在不同类型的无线通信（见图 6.2）：

1）V2V 通信：V2V 通信由不同 OBU 之间的数据交换和通信组成。

2）车辆到路侧（V2R）通信：这些通信与路侧通信基础设施有关。

3）V2I 通信：V2I 通信由一个 OBU 和一个 RSU 之间的数据交换组成，由一个 OBU 中继（V2V2R）。V2I 通信也由两个 OBU 之间的数据交换组成，由一个 RSU 中继（V2R2V）。

4）车辆内通信：这些通信是在一个车辆内车载设备和传感器间发生的。

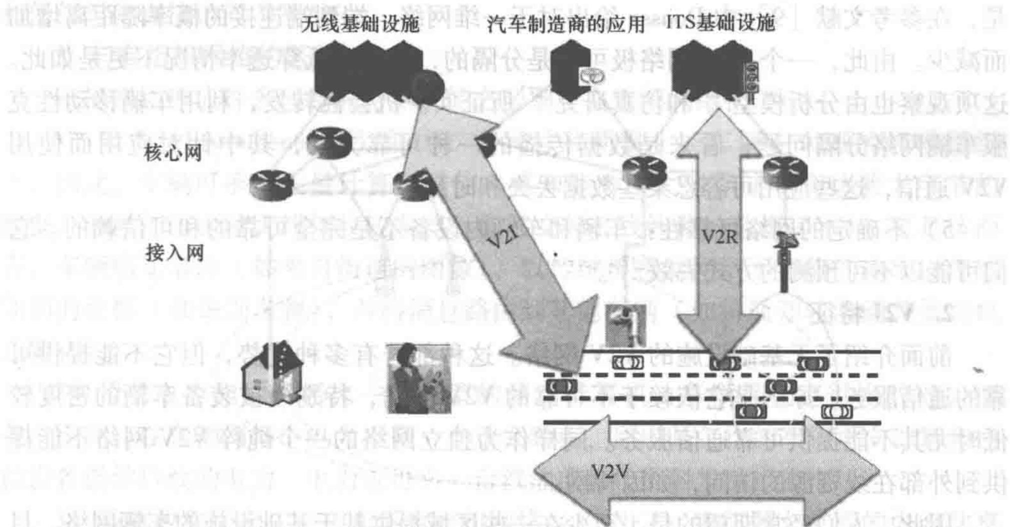


图6.2　车辆通信

在本章中，主要关注车辆到车辆和车辆到基础设施这两种通信方式。接下来，

将说明V2V和V2I通信的基本特征。

1. V2V特征

一个V2V网络是一个无基础设施的网络，仅由装备的车辆所组成。典型情况下，车辆装备有短距离通信设备，并可与其无线范围内的其他车辆交换信息，得以创建自组织无线网络，由该网络传播信息。

V2V通信适合主动安全和实时状况感知以及其他应用。因此，需要它们是快速的、可靠的和简单的。

V2V部署提供低成本和容易部署的优势，对一些局部化的应用（如协作驾驶）而言是必要的。一个V2V网络是一种特殊类型的自组织网络，展示出一些独特的特点[13,23,26]：

1）可预测的高移动性：车辆经常以高速运动，但在路上其移动性是相当规则的和可预测的。事实上，车辆运动对道路是空间受限的，车辆操作受限于车辆性能限制和交通法规，如最大和最低速度。

2）动态的但地理上受限的拓扑：虽然由于车辆的高移动性，车辆之间的互联可快速变化，但道路网络经常将通信网络拓扑限制到一维空间。一个V2V网络可被设想为重叠在道路网络之上。即使在道路处在附近范围时，一般而言，障碍物（如建筑物）使无线信号不能在道路之间传播，例外情况是十字路口附近。

3）可能的大型规模：在参考文献中研究的多数自组织网络，通常假定一个有限的范围，与此不同的是，从原理上而言，V2V网络可在整个道路网络上扩展。

4）分隔的网络：经常在自组织联网研究中，隐性地假定端到端连接性。但是，在参考文献［9］中Dousse给出对于一维网络，端到端连接的概率随距离增加而减少。由此，一个V2V网络极可能是分隔的，特别在低穿透率情况下更是如此。这项观察也由分析模型[26]和仿真研究[27]所证实。机会性转发，利用车辆移动性克服车辆网络分隔问题，看来是数据传播的一种可靠方法，其中针对应用而使用V2V通信，这些应用可容忍某些数据丢失和时延。

5）不确定的网络可靠性：车辆和车辆内设备不是完全可靠的和可信赖的。它们可能以不可预测的方式失效[17]。

2. V2I特征

前面介绍了无基础设施的V2V网络。这种部署有多种优势，但它不能提供可靠的通信服务，原因是它依赖于不可靠的V2V通信，特别当被装备车辆的密度较低时尤其不能提供可靠通信服务。同样作为独立网络的一个纯粹V2V网络不能提供到外部在线资源的访问，如因特网。

因此，人们经常期望的是，至少在一些区域提供基于基础设施的车辆网络，目的是提供可靠的宽带通信服务和访问没有驻留在车辆上的在线资源与本地服务（如交通信息、旅游信息）。

V2I基础设施提供两种类型的访问：功能特定的端口和通信端口。针对特定任

务，车辆与前者通信。例子有支持无线的十字路口控制器［支持信号抢占（为紧急车辆让路）］或信号优先级［优先处理大众运输车辆］、坡道距离控制器以及收费和停车支付采集器。

通信端口［如接入点（AP）和广域无线网络（WWAN）基站（BS）］代表另一种类型的访问，提供网络访问能力。

从网络基础设施部署的角度看，V2I 通信形成许多独特的特点。

1）车辆分布：车辆分布值得特别关注。常规移动用户经常被假定集中在某些热点区域，如建筑物、机场、咖啡馆等。但是，车辆经常是大范围分布的，相比其他道路（邻接街道），其中一些道路［如免收费公路］可能有较高的车辆集中度。在任何特定的道路段，由于事故，车辆分布可能发生急剧变化；在其他情形中，车辆集中度可能多少是比较可预测的，像在高峰时间或由于道路施工造成的拥塞中的情形。

2）无线基础设施部署：在财务成本、地理范围和要支持的用户数方面，车辆的无线基础设施的部署给出前所未有的挑战。从商业角度看，这样的一项基础设施也许可作为支付用户的一个优惠服务加以提供，就像如今提供的蜂窝和多项无线局域网（WLAN）一样。另外，服务也许可由政府部署，如用于交通监测和管理目的或经济发展目的。

3）无线技术：基础设施可利用各种无线技术（如 WWAN 和 WLAN）一起工作在无缝方式下。在城市网络中，如 WWAN BS 和 WLAN AP 都通过有线链路或固定宽带无线链路连接到一个骨干，骨干自己也被连接到因特网。用户可在任何地方和任何时间直接访问 WWAN。

3. 与自组织网络的区别

车辆网络形成许多独特的特点，使之区别于自组织网络：

1）装备（Instrumentation）能力：在尺寸和质量上，车辆在大型的规模量级上。因此，车辆可承担大量计算、通信、感知能力、大型存储（高达数太字节的数据）和功能强大的无线接收转发器（能够交付有线线路的数据速率）。更具体而言，车辆感知事件（如来自街道的图像），以传统传感器网络不可能的速率处理感知到的数据（如识别车牌），并将消息路由到其他车辆（如将通知转发到其他司机或警官）。

2）设备电力问题：相比一台典型的移动计算机，车辆具有高得多的电力储备。因此，在车辆网络中电力不是一个重大约束，原因是工作的车辆可向计算和通信设备提供持续的电力。电力也可由一台汽油或替代燃料发动机充电。

3）数据收集平台：一个车辆自组织网络（VANET）为部署大范围的内容共享应用（对等应用）提供了机会。这个问题构成移动数据收集的一个理想平台，特别在监测城市环境的语境（即车辆传感器网络）中更是如此。

4）车辆速率：车辆是以非常高的速度在运行的。结果是，持续的、一致的

V2V 通信是难以维持的。尽管如此，车辆运动的已有统计信息表明，在通勤时间期间一起运行的趋势或交通模式可帮助维持移动车辆群间的连接能力。

5）车辆基础设施邻域：车辆总是距离基础设施（Wi-Fi、蜂窝、卫星等）有几跳远。在这个语境中，为了容易地访问因特网，必须考虑网络协议和应用设计。

6）短的路由寿命：在一个高度运动的环境中，路由发现和路由维护的问题就出现了。两辆车之间的平均链路寿命在几秒到 10s 范围。当设计一个车辆网络路由协议时，应该考虑这个问题。

7）寻址：在车辆网络中，寻址是不同的。通信使用地理位置进行寻址和分组转发。在一个特定地理区域中与车辆的信息交换（可能距离信息源较远）要求可靠的和可扩展的通信能力。我们称这些能力为地理寻址和路由（地理联网）。

8）不同的应用：除了信息娱乐和舒适度应用外，车辆网络提供了新的安全应用，这在下一节重点讨论。

6.4 车辆应用

车辆网络为各种应用和服务打开了大门，范围从自动高速系统到分布式乘客电话会议。这些应用可被分类为安全应用和非安全（舒适或方便性）应用。

安全应用[17]吸引了人们相当的关注，原因是它们直接与最小化道路上的事故数有关。另外，非安全应用可包括旅途规划实时道路交通估计、高速收费、协同探险、信息检索和娱乐应用。

由于安全应用和非安全应用的不同特征，它们有不同的需求。这些需求与连接能力、作用距离、模式、延迟/抖动、分组交付率、数据尺寸/连接时长、服务交付、安全隐私有关，如表 6.1 所示。接下来，将详细描述安全应用和非安全应用的特征。

表 6.1　车辆应用需求（参见美国丰田信息技术中心的网页）

	安全应用	非安全应用
连接能力	先验式的（总是在线）	应需/事务
作用距离	局部	远距离
模式	地理广播/组播	单播/组播
延迟/抖动	紧急/甚低	各种
分组交付率	高	各种
数据尺寸/连接时长	小/短	大型/各种
服务交付	所有邻居/一些	位置感知/广域
安全	要求	要求
隐私	要求	要求

6.4.1 安全相关的应用

车辆网络的最紧迫应用与安全特征有关，并应在所有车辆上提供。安全相关的应用通常要求直接通信，这是由于其时延关键的性质决定的。一个这样的应用将是紧急通知，如紧急刹车告警。在一次事故（气囊触发事件）或突然的硬性碰撞的情形中，一条通知被发送到后跟的汽车。那个信息也可由以相反方向运动的汽车传播，由此，也许被传递到发生事故的车辆。

安全相关的应用可被分组为三个主要类别：信息型的、辅助型的和告警型的。

1. 信息型应用

使用内部和外部源，司机信息应用提供有关车辆周边环境和车辆本身的信息。

传播到道路司机处的信息，可帮助他们适应当前道路状况。一个这样的应用可以是有关速度限制或工作区的信息传播。

2. 辅助型应用

这些应用提供协作的司机和变道服务。

1）一项高级辅助服务是协作司机辅助系统，该系统利用了汽车间传感器数据或其他状态信息的交换。协作驾驶系统要求 V2V 和 V2I 协作。这些应用的例子有自适应巡航控制、成组和自适应驾驶。

协作驾驶系统辅助司机保持车辆之间的安全的时间——前行距离，以便确保紧急刹车不会导致汽车之间的碰撞。通过考虑变化的环境状况、车辆动态和安全，时间间隔计算系统调整一辆车的时间间隔距离。更具体而言，如果到前面汽车的距离改变，则通过加速或刹车，协作司机系统必须相应地做出响应。

2）变更车道辅助（LCA）：这项应用辅助司机变更车道[20]。相对车道边界，该系统监测车辆的位置。如果发起一次车道变更操作，且该系统在邻接车道监测到一辆车，则系统提醒司机。针对信息交换，使用基于 VANET 的无线技术，则这种做法为司机提供一项附加工具，确定交通条件是否允许开始一项接管操作。这项应用辅助司机选择接管的最佳时刻，并影响司机的行为朝着改进驾驶技能发展，由此降低道路事故。

3. 告警型应用

这些应用提供有关未来危险道路状况、障碍、易出错的司机和有优先权的车辆（紧急车辆）的信息。内部传感器、其他车辆和基础设施提供有用信息，这些信息由碰撞后告警应用使用。基本想法是拓宽他或她（司机）的视野外的感知范围，并进一步以自动辅助应用辅助司机的操作。

在这个分类内，提供几项服务：

1）前方碰撞告警（FCW）：FCW 系统检测一次即将发生的碰撞。取决于系统，它可能警告司机、提前接管刹车、为额外支撑而抬高座位、将乘客座位移动到

一个较佳位置、为缓解冲击而折叠后头枕、收紧安全带、去除额外的松弛距离，并自动地应用部分或全量刹车，以便最小化碰撞严重程度。

2）电子紧急刹车灯（EEBL）：EEBL 增强司机的可视能力，方法是通过车辆间的无线链路传播告警消息，并以最小的延迟将有关急迫状况的告警通知告知处于危险境地的司机[28]。EEBL 应用也许不仅提高一条硬刹车消息的告警范围，而且提供重要信息，如加速/减速速率。

3）十字路口碰撞告警：为避免十字路口碰撞，需要提前向司机提供十字路口附近的必要信息[6-8,10]。例如，当接近或通过十字路口时，一名司机应该被通知即将发生的碰撞。协作的 V2I 技术在避免十字路口处的碰撞方面为司机提供辅助，这种碰撞是由于走神、错误感知、遮挡的视野或醉酒导致的。这些类型的系统由如下车辆组成，它们持续地将信息中继到位于接近十字路口处的一个灯标。

6.4.2 非安全（便利性、舒适度）应用

这些应用的一般目的是改善乘客的舒适度和交通效率。舒适度应用的重要特征是，它们不应干扰安全应用。在这个语境中，交通优先权和独立物理信道的使用是一种可行的解决方案。

舒适度应用有车内娱乐、车载共享、交通管理和货物应用。

1. 车内娱乐应用

这些应用为乘客提供音频和视频数据，这些数据是从其他车辆或基础设施得到的。所有种类的应用［它们可能运行在传输控制协议/因特网协议（TCP）/IP 栈之上］，也许可适用这里的情况，如在线游戏或即时消息传递。

另一项应用是从商业车辆和路侧基础设施接收到有关其商务的数据（无线广告）。企业（购物中心、快餐、加油站、酒店）可建立静态网关，将营销数据传递给路过的潜在客户。

2. 车载共享应用

车载共享应用在车辆上分派数据或计算。它们依赖于车辆间的共享系统。

一项有趣的应用是，使用分布式的车辆计算资源，实时地测量道路总的碳排放。无论何时碳排放达到一个临界阈值时，车辆可调整其行为，降低污染水平，做法是关闭它们的适应系统（Acclimatization System）、降低速度或在交通堵塞时关闭发动机。

3. 交通管理应用

高速塞车正在向司机施加一项不可容忍的负担。因为当对出行的要求超过高速容量时就发生塞车，所以降低塞车的一种听起来可行的方法，将涉及各种政策的混合使用，来影响要求和容量，这取决于当地环境和优先级[14,18,21,22]。这些政策之一是实施交通管理应用。

交通管理应用提供各种服务，其中有[24,25]交通管理中心、电子道路收费

(ETC) 系统和智能交通信号。

(1) 交通管理中心

这被用于 ITS，引导车辆交通。它提供如下服务：

1) 交通报告，为道路用户提供建议。

2) 导航系统，帮助司机定位最优路线、宾馆、餐馆等。这些服务是基于位置的，并基于车辆地理位置显示信息。这些系统中的软件执行直接受到外部环境的影响。

3) 交通统计器，提供实时的交通统计信息。

4) 融合指标道路交通监测，提供有关高速入口匝道使用情况的信息。

5) 停车指南和信息系统，向汽车驾驶员提供免费停车的动态建议。

(2) 电子道路收费 (ETC) 系统

ETC 系统一直被看作对新基础设施融资和改善交通流的一项有效方式。ETC 也可节省道路出行人员的时间，消除沮丧感，使它们能够不停车地通过收费区。当一辆汽车通过一个收费点时，一个路侧天线与安装在汽车仪表板内部或汽车风挡玻璃后面的 OBU 交互通信。当车辆通过收费区时，它们被自动地计费，这就提供了吞吐量并最小化了时延。

(3) 智能交通信号

将分布式控制系统应用到交通管理，被称作"智能信号"[12]。它基于通过以太网连接的空间分布的微处理器。微处理器将复杂数据传递给交通控制器。为得到改善的服务质量，系统对人们和车辆的个体需求做出响应。为取得较高的有效性，交通控制器设备为车辆司机显示分钟级的准确信息。

使用智能信号，交通控制器可识别没有以典型移动速度运动的步行者，之后通过延长信号灯的时间做出响应，以便为步行者通过十字路口提供附加的时间。信号灯的时长也可由微控制器修改，以便处理具有长刹车距离的车辆。

4. 货物应用

1) 车辆注册、检查和证书：车辆检查是最重要的安保和安全措施之一，用来防止在途的事故，并控制商品/个人运输的合法性[11]。拦停一辆车，验证司机驾照的有效性，或检查车辆或旅行文档 [如危险商品进入一个集装箱码头 (装卸区) 的安全卡]，或在车辆进入一个道路基础设施之前检查车辆的物理状态，这些活动是车辆检查的典型例子。无线车载网络支持车辆和道路基础设施 [如道路、隧道、集装箱码头 (装卸区)] 之间的数字服务交换，并使大量重要的车辆数据 (如发动机状态、胎压、货物文档) 直接可用于基础设施信息系统应用。

2) 货物监测和跟踪：对于从市内到室外和从仓库到集装箱的运输中转中，车载环境的无线接入，在货物层面填补了无缝和连续跟踪的鸿沟。车载网络将形成一个跟踪系统，它支持连续的和泛在的货物层面的监测。

6.5 车载标准

几项无线网络技术将为 ITS 通信铺平道路。虽然 IEEE 802.11p 是 V2V 通信物理和媒介访问控制（MAC）层的建议标准，但高速分组接入（HSPA）、长期演进（LTE）和 IEEE 全球微波接入互操作性（WiMAX）/802.16e 也鼓吹可用于 V2I 通信。

在一个区域中可共存多项无线技术。农村和城市区域可部署不同的网络架构。在城区，诸如 V2I 通信的无线基础设施，提供几乎泛在的连接能力，Wi-Fi 部署则持续地变得范围越来越大。V2V 通信也可用于直接的车辆间信息交换。

在农村地区，也许比较经济的做法是依赖 V2V 通信，并在某些热点或特别关注的其他区域所放置的有限基础设施加以补充。

接下来，将给出为 V2V 通信建议的标准，即 CALM、C2C-CC 和 WAVE。

6.5.1 陆地移动的通信接入

1. CALM 概念和优势

CALM 是针对使用几种媒介中一种或多种媒介的无线数字数据通信的、空中接口协议和参数的一个标准化集合。

CALM 支持未来通信技术、联网协议和高层协议，以便支持高效的 ITS 通信服务和应用。CALM 提供一个通信子系统，该系统：

1）无论一个车辆存在于一个交通状况的哪里和何时处于其中，它总是可用的。

2）以一种透明方式可在车辆-车辆和车辆-路侧间通信。

3）使应用从需要知道通信设置和管理的需要中解脱出来。

4）使用现代因特网技术和标准，解决全球可用性。

5）提供与数据速度、通信距离、成本和许多其他参数有关的不同可能性的一个范围。

CALM 规范和标准不是有关实现设备的一个物理部分的，CALM 实际上是一组协议、规程和管理过程。物理设备的实现是商业相关过程的一个函数。

2. CALM 通信模式

CALM 通信服务包括如下通信模式：

1）车辆-车辆：这是一个低延迟对等网络，具有这样的能力，携带诸如碰撞避免的安全相关数据和诸如自组织网络连接多个车辆的其他车辆-车辆服务。

2）车辆-路侧：采用这种类型的通信，路侧站没有连接到一项基础设施。但是，它也许被连接到一个十字路口周边 ITS 站组成的一个局部网络。

3）车辆-基础设施：自动地协商多点通信参数，可由路侧或车辆发起后续通

信。路侧站被连接到一项基础设施，如因特网或其他网络。

4）基础设施-基础设施/路侧-路侧：通信系统也可能被用来连接固定点，其中不期望使用传统的线缆方式。

3. CALM 媒介

CALM 媒介被定义为

1）5GHz 无线局域网（LAN）系统基于 IEEE 802.11 常规 Wi-Fi 和新的 CALM M5/802.11p 模式。

2）蜂窝系统，GSM/HSDSC/通用分组无线服务（GPRS）和第三代（3G）通用移动电信系统（UMTS）。

3）60GHz 毫米波系统。

4）红外通信。

5）一个汇聚层支持专用的短距离通信（DSRC）、广播和定位。

CALM 架构和标准的原则是在如下原则下可预测的，即最佳使用可用的资源：CALM 使用最优存在于任何特定位置的无线电信媒介，并当必要时，具有切换到一种不同媒介的能力。

CALM 将同时支持多种类型的应用和多种类型的媒介。但是，为支持所有可能的媒介，对实现的设备却没有要求：要支持何种媒介的选择，将是设备或车辆制造商的决定，也取决于可用的媒介选项，国与国是不同的，从一个地方到另一个地方也可能是不同的。

采用 CALM，并不意味着实现它的所有可能性：CALM 支持各组件在支持可用媒介的任何地方无缝地工作。

4. CALM 标准

CALM 标准正在由 ISO TC204 工作组 16[16,19] 开发。在如下标准中描述了 CALM：

1）ISO 21217：CALM 架构。

2）ISO 24102：CALM 管理。

3）ISO 21218：CALM CI 服务 AP。

4）ISO 21210：CALM 因特网协议版本 6（IPv6）联网。

5）ISO 29281：CALM 非 IP 联网。

6）ISO 24101：CALM 应用管理。

5. CALM 协议栈

ISO 21217 描述通用架构框架，围绕该框架，实例化符合 CALM 的通信实体（称作 ITS 站），并提供 CALM 国际标准族使用的架构参考，包括低层服务 AP、网络协议规范（IPv6 联网和非 IP 联网）和 ITS 站管理规范。

应用层、网络和传输层、接入层将构成 ITS 主机架构，如图 6.3 所示。该标准规范了一个通用架构、网络协议以及用于有线和无线通信（使用各种接入技术）

的通信接口定义。

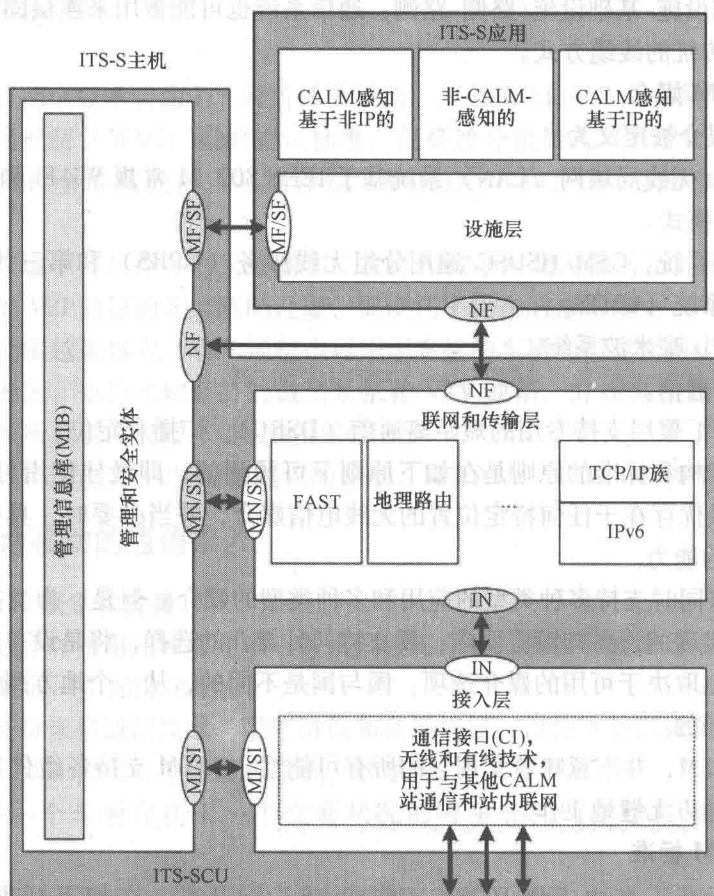


图 6.3　CALM 标准化的协议层[19]

设计接入技术是为了在移动站之间、移动站和固定站之间以及 ITS 扇区中固定站之间提供广播、单播和组播通信。

可以想象的是，CALM 将包括现有通信技术和 CALM 特定的通信技术。在这个语境中，CALM 感知的基于非 IP 的和基于 IP 的应用以及不感知 CALM 的应用，将共存并可用于不同的 ITS 主机（见图 6.3）。

CALM 概念优于传统系统的一项基础优势是，各应用是从接入技术（提供无线连接）和网络［将信息从源传输到目的地（多个）］抽象出来的。这意味着 ITS 站不会受限于单项接入技术和联网协议，并可实现所支持的任意一种技术和协议。结果，ITS 站管理可最优使用所有这些资源。

6. CALM 切换支持

在 ITS 中，对于诸如安全，交通信息和管理，为旅游信息、娱乐而将视频下载

到移动站，以及导航系统更新等目的，要求大体量的数据。为支持这样的服务，移动站需要能够在较长距离上与固定站通信，系统必须能够从一个固定站将会话切换到另一个固定站。

由此，明确地设计 CALM 国际标准族，支持准连续的通信、延长时段的通信、短消息和具有严格时间约束的高优先级会话。

CALM 概念的本质特征之一是支持媒介无关切换（MIH）的能力，也称作异构切换，这是在由 CALM 所支持的各种接入技术之间切换的，如蜂窝、卫星、微波、移动无线宽带、红外和 DSRC。

采用这种灵活性，符合 CALM 的系统提供针对消息交付使用最适合接入技术的能力。在做出判断的过程中，选择规则包括用户偏好和接入技术能力，这样的判断是就一个特定会话使用哪种接入技术以及何时在接入技术之间切换或在相同接入技术上服务提供商之间的切换做出的。

为利用这种灵活性，符合 CALM 的系统提供支持不同类型切换的能力，包括涉及改变通信接口的那些切换（可能或不可能涉及改变接入技术，原因是 ITS 站可能有使用相同接入技术的多个通信接口）、涉及网络（用来提供连接能力）重新配置或改变的那些切换、涉及通信接口和网络重新配置改变的那些切换。

6.5.2　汽车到汽车（汽车间）通信联盟

1. C2C-CC 概念

C2C-CC 的目标是标准化车辆及其环境之间无线通信的接口和协议，目的是使不同制造商的车辆可互操作，也使它们可与 RSU 通信[1]。

C2C 系统提供如下顶级特征：

1）车辆之间以及车辆和 RSU 之间的自动快速数据传输。

2）交通信息、危险告警和娱乐数据的传输。

3）在不需要预安装的网络基础设施条件下，支持自组织服务。

4）在短距离 WLAN 技术上传输，无传输成本。

自组织 C2C 支持车辆协作，方法是连接分布于多个车辆间的个体信息。如此形成的 VANET 工作起来像一个新的传感器，增加了司机对热点的感知距离，司机和板上传感器系统以其他方式都是看不到的。

C2C 系统以电子方式扩展了司机的视野，并完全支持新的安全功能。C2C 通信形成去中心化主动安全应用的一个良好基础，因此将降低事故及其严重性。除了主动安全功能外，它们包括主动交通管理应用，并有助于提高交通流。

2. C2C-CC 域

C2C 通信系统的架构如图 6.4 所示。它由三个不同的域组成：车内、自组织和基础设施域。

(1) 车内域

车内域指逻辑上由一个OBU和多个应用单元（AU）组成的一个网络。一个AU典型地是一台专用设备，它执行单项应用或一组应用，并利用OBU通信能力。一个AU可以是一辆车的一个集成部分，并被永久地连接到一个OBU。它也可以是一台便携设备，可动态地连接到一个OBU（并从此处断开）。

(2) 自组织域

自组织域或VANET，由装备有OBU的车辆和沿路的静态单元（称作RSU）组成。一个OBU至少装备有一台（短距离）无线通信设备（专用于道路安全），并可能装备有其他可选的通信设备。各OBU形成一个MANET，这以一种全分布式方式支持节点间的通信，而不需要一个中心式协同实例。

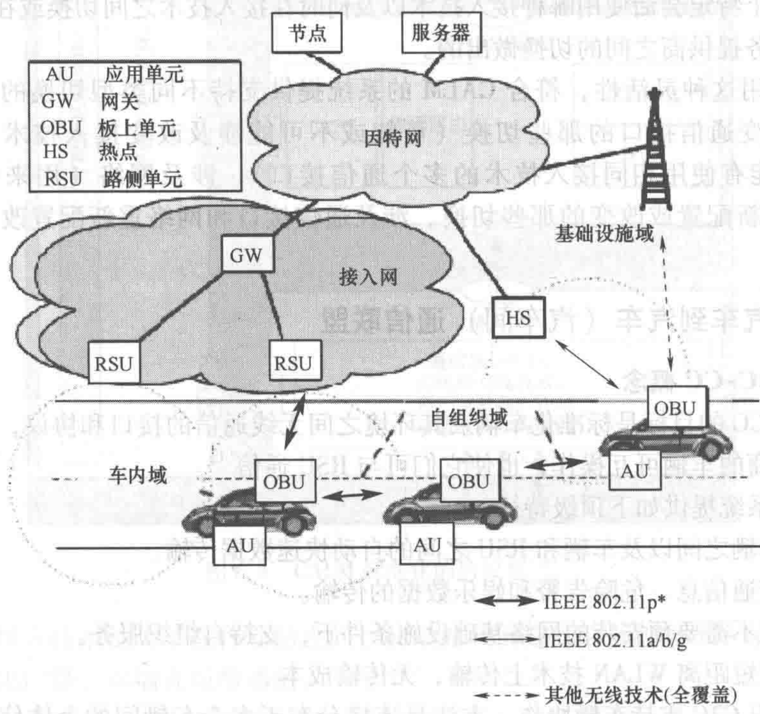


图6.4　C2C-CC域[1]

一个RSU的主要角色是改善道路安全，方法是执行特殊应用，并在自组织域中发送、接收或转发数据，扩展自组织网络的覆盖范围。一个RSU可连接到一个基础设施网络，该网络接下来可连接到因特网。

一个RSU的主要功能有：

1）当OBU进入RSU的通信范围时，通过将信息重新分发到该OBU，就扩展了一个自组织网络的通信范围。这项功能包括这样的情形，其中一个RSU以车辆的无线多跳链方式直接转发数据（见图6.5）。

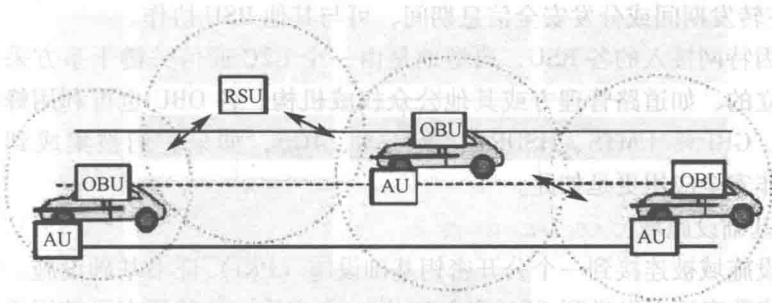


图 6.5　通过转发数据[1]，一个 RSU 扩展一个 OBU 的通信范围

2）可能运行安全应用，如针对 V2I 告警［例如，矮桥告警、工作区告警］、十字路口控制器或虚拟交通信号，并分别作为信息源和接收方（见图 6.6）。

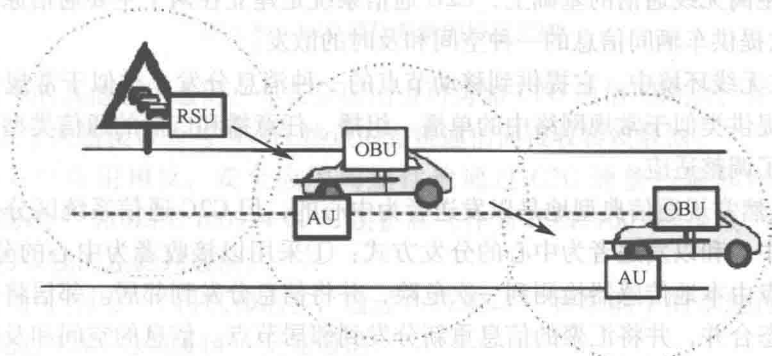


图 6.6　一个 RSU 作为信息源[1]

3）可能向各 OBU 提供因特网连接能力（见图 6.7）。

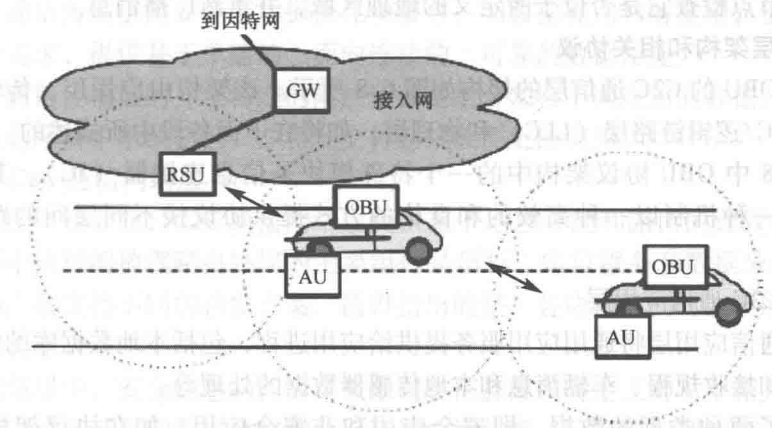


图 6.7　一个 RSU 提供因特网接入[1]

4）在转发期间或分发安全信息期间，可与其他 RSU 协作。

用于因特网接入的各 RSU，典型地是由一个 C2C 通信关键干系方采用一个受控过程建立的，如道路管理方或其他公众权威机构。各 OBU 也可利用蜂窝无线网络（GSM、GRPS、UMTS、HSDPA、WiMAX、4G），如果它们被集成到 OBU 中，特别对于非安全应用更是如此。

（3）基础设施域

基础设施域被连接到一个公开密钥基础设施（PKI）证书基础设施。证书权威（CA）是将数字证书发送到 OBU 和 RSU 的一个实体。这些证书可被用于节点间的通信之中，验证安全证书是否属于某个节点。其使用意图用于一项整体安全策略，这超出了本书的范围。

3. 基本通信原理

在短距离无线通信的基础上，C2C 通信系统是建立在两个主要通信原理上的：

1）它提供车辆间信息的一种空间和及时的散发。

2）在无线环境中，它提供到移动节点的一种消息分发，类似于常规分组交付网络，并提供类似于常规网络中的单播、组播、任意播和广播的通信类型，但对车载环境做了调整适应。

3）虽然常规通信典型地是以发送者为中心的，但 C2C 通信系统区分信息的以接收者为中心和以发送者为中心的分发方式：①采用以接收者为中心的分发方式，一个源节点由本地传感器检测到一次危险，并将信息分发到邻居。邻居将信息与本地信息状态合并，并将汇聚的信息重新分发到邻居节点。信息的空间和及时分发是受到接收节点控制的，接收节点作为一个转发器：在接收到信息时，它确定信息对其邻居的相关性，并判定是否应该重新分发该信息。②采用以发送者为中心的分发方式，一个源节点定义一个地理区域，并将信息转发到所有邻居。在接收到信息时，邻居节点检查它是否位于所定义的地理区域，并重新广播消息。

4. 分层架构和相关协议

一个 OBU 的 C2C 通信层的架构如图 6.8 所示。该架构由应用层、传输层、网络层、MAC/逻辑链路层（LLC）和物理层，如将在下面各段中所描述的。

图 6.8 中 OBU 协议架构中的一个特殊模块是信息连接器（IC）。其主要任务是通过一种机制以一种高效的和良构的方式提供协议栈不同层间的跨层数据交换。

（1）C2C 通信应用层

C2C 通信应用层将通用应用服务提供给应用进程，包括本地数据库的维护、消息的发送和接收规程、车辆消息和本地传感器数据的处理等。

定义了两种类型的数据，即安全应用和非安全应用。如在协议架构（见图 6.8）中看到的，非安全应用使用带有 IPv6 之上 TCP 和用户数据报协议（UDP）（或另一种传输协议）的传统协议栈，并可访问无线多跳通信，与车辆、RSU 或因

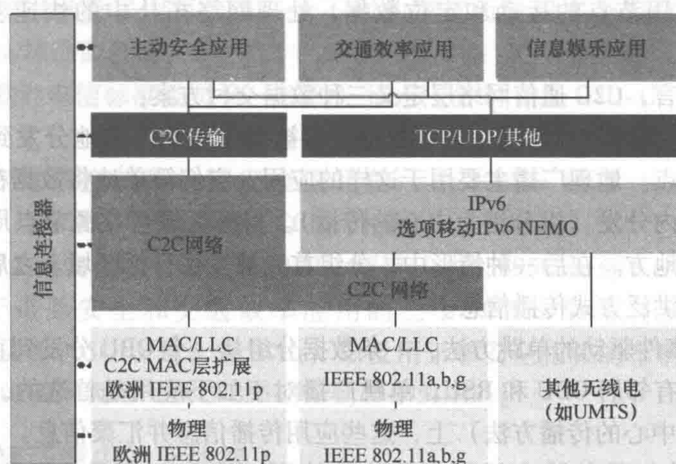


图6.8 C2C通信系统的协议架构[1]

特网节点中的其他应用通信。非安全应用也可旁路C2C通信网络层，并通过IEEE 802.11a/b/g网络接口为与Wi-Fi热点的直接通信而接收转发数据。

与非安全应用相反，安全应用周期性地通过C2C通信传输和网络层以及802.11p物理[15]和IEEE 1609.4 MAC层扩展（作为IEEE 1609标准族[2-5]的组成部分），与协议栈的左侧列通信。

各应用可与用户（司机和乘客，通过人机接口）和车辆中的本地传感器数据（典型地通过CAN总线接口）交互通信。

C2C通信基本系统由一组应用组成，这些应用在每个车辆中必须实现。但是，可安装和执行其他应用（扩展系统）。

（2）C2C通信传输层

C2C通信传输层向安全应用提供几项服务，如数据复用和解复用，也可依据安全应用的需求，提供基于单播的、面向连接的、可靠的数据传递。

C2C通信传输层的一项特定的附加任务是组合来自不同应用的数据，目的是在单条分组的净荷中承载它们，并将其交付到接收侧上的应用。

（3）C2C通信网络层

在无线电层上面，C2C通信网络层提供基于地理寻址和路由的无线多跳通信。在网络层中执行的地理路由协议的主要组件是信标、定位服务和数据分组的转发。为单播和广播支持不同的转发方案。值得指出的是，各应用可同时或顺序地使用这两种通信类型。

在网络层中，安全信息的传播可被限制在信息源发者定义的一个相关区域。这可以如下方式做到，即数据分组朝目标区域中继传播，一旦它们到达地理上的目标区域，它们就被高效地传播到目标区域内的所有车辆。

单播数据分组从源通过多跳通信被转发到目的地。通过VANET定义路径的路

由算法，可使用节点的运动和定位数据，处理网络拓扑中的快速变化［“地理单播”］。

更具体而言，C2C 通信网络层定义三种数据交付方案：

1）采用事件驱动的地理广播，数据分组被高效地和可靠地分发到一个地理区域内的所有节点。地理广播主要用于这样的应用，它们简单地将数据在一个明确定义的地理区域内分发（以分组为中心的传播）。目标区域可在源节点周边，但它可位于比较远的地方。在后一种情形中，分组首先被发往目标区域。之后，在目标区域，该分组以洪泛方式传播信息。

2）采用事件驱动的单跳方法，一条数据分组从一个 OBU 分发到直接无线通信范围内它的所有邻居 OBU 和 RSU。单跳广播对于如下应用是首选的，在每个无线跳（以信息为中心的传播方法）上，这些应用传播信息并汇聚信息。

3）信标分组是单跳广播的一种特殊情形，是由 C2C 通信网络层周期性发送的。

（4）C2C 通信 MAC/LLC 层

下面描述被选中的几个设计原则，它们被识别为 C2C-CC 中的基础。

C2C-CC MAC 层基于 IEEE 802.11 MAC 协议（在参考文献［15］中规范），但在服务方面有多项简化，在跨层集成方面有一些增强。被采用的 MAC 算法是标准的带有冲突避免的载波侦听多路访问（CSMA/CA）。

C2C 通信 MAC 层定义了单个自组织网络，其中遵循 C2C-CC 标准的所有节点都是成员，而不需要任何关联规程。

就拥塞控制而言，C2C-CC 识别出如下必要的特征，但它们却没有被包括在 802.11 标准中：

1）MAC 层应该向高层提供有关当前所估计信道负载的信息。依据这个信息，上层应用不同战略以防止媒介拥塞（例如，取决于优先级，应用判定它们是否可进行传输）。

2）LLC 子层应该向网络层提供每分组参数控制，特别就传输功率而言更应如此。

3）要求用于信道观测的一个客户端/服务器接口以及 MAC 层和所有高层之间的控制命令。

4）按照应用所指定的，依据消息的优先级，MAC 层应该实现一种区分性的排队方案。

（5）C2C 通信物理层

C2C-CC 主要区分三种类型的无线电（射频）无线技术：IEEE 802.11p、基于 IEEE 802.11a/b/g/n 的常规无线 LAN 技术和其他无线电技术（像 GPRS 或 UMTS）。

从架构观点看，C2C-CC 考虑即将到来的 IEEE 802.11p（车辆环境中的无线接入）无线电技术，带有适应于车辆环境的修改和修正，像不使用关联/认证、特定

传输功率的使用以及采用每信道一个不同的带宽而使用多个信道，这些是相比 IEEE 802.11a 标准而言的。

已经在欧洲电信标准委员会（ETSI）请求了针对专用 C2C-CC 信道的如下频带分配：

1）用于网络控制和关键安全应用的一个 10MHz 频带（从 5.885GHz 到 5.895GHz），与 WAVE 控制信道（CCH）相同。

2）用于关键安全应用的一个 10MHz 频带（从 5.895GHz 到 5.905GHz）。

3）用于道路安全和交通效率应用的三个 10MHz 频带（从 5.875GHz 到 5.885GHz、从 5.905GHz 到 5.925GHz）以及用于非安全相关的车辆到路侧和 C2C 应用的两个 10MHz 频带（从 5.855GHz 到 5.875GHz）。

6.5.3　车辆环境中的无线接入

1. WAVE 概念

平行于 CALM 和 C2C-CC 标准，WAVE 是一种无线电通信系统，意图为运输提供互操作的无线联网服务。这些服务包括由美国国家智能运输系统架构（NITSA）针对 DSRC 识别出的那些服务以及在架构中没有特别识别出的许多其他服务。

该系统支持 V2V 和 V2R 或 V2I 通信，一般是在小于 1000m 的视距距离上传输的，其中车辆可能以高达 140km/h 的速度移动。

2. WAVE 标准

在标准中定义的 WAVE 协议栈的各组件如图 6.9 所示。

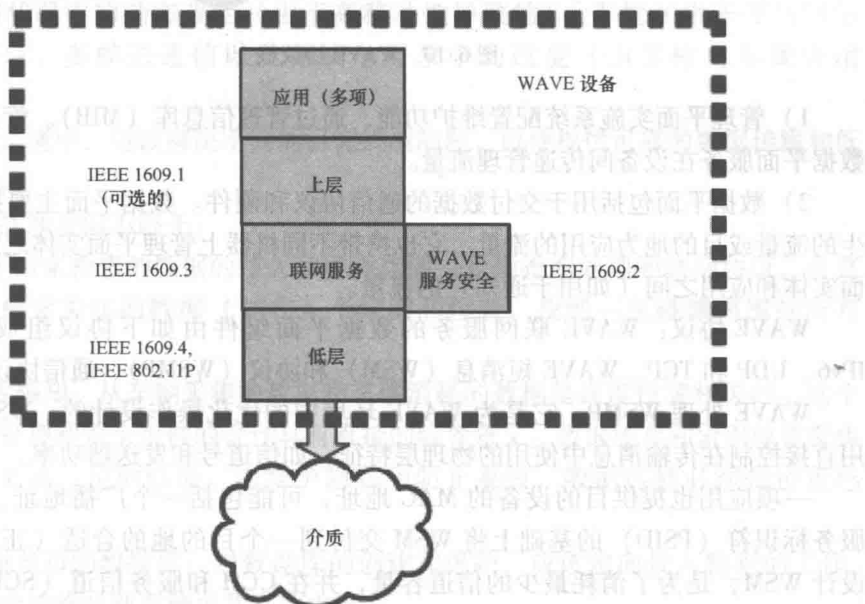


图 6.9　WAVE 标准[4]

物理层和 MAC 层使用 IEEE 802.11p[15] 和 IEEE 1609.4[5] 标准的各组成单元。在 IEEE 1609.3[4] 中定义了联网服务。另外，文档“IEEE Std 1609.2”[3] 为 WAVE 联网栈和意图运行在那个协议栈的应用规范了安全服务。各服务包括使用另一方公开密钥和非匿名认证的加密（服务）。

IEEE 1609.1[2] 定义了一项应用，即资源管理器，它为通信使用协议栈。

3. WAVE 协议栈

从 WAVE 联网服务的角度看，整个 WAVE 协议栈如图 6.10 所示。该栈由如下部分组成：

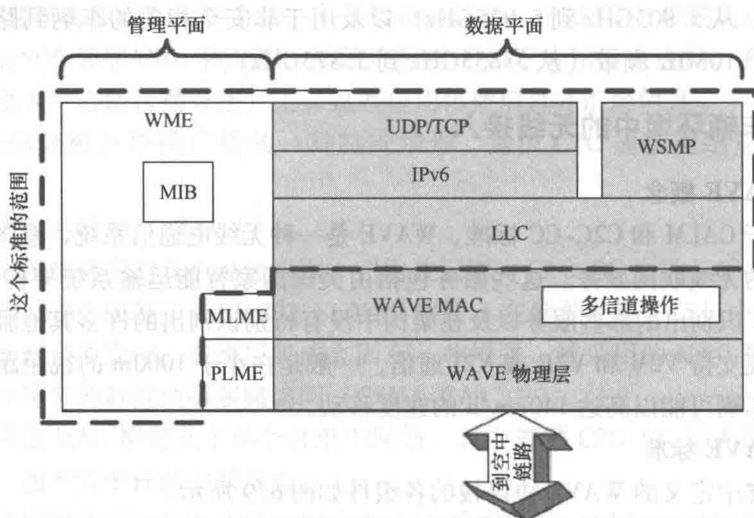


图 6.10　WAVE 协议栈[4]

1）管理平面实施系统配置维护功能，通过管理信息库（MIB）。管理功能利用数据平面服务在设备间传递管理流量。

2）数据平面包括用于交付数据的通信协议和硬件。数据平面主要携带应用产生的流量或目的地为应用的流量。它也携带不同机器上管理平面实体之间或管理平面实体和应用之间（如用于通知）的流量。

WAVE 协议：WAVE 联网服务的数据平面组件由如下协议组成，即 LLC、IPv6、UDP 和 TCP、WAVE 短消息（WSM）和协议（WSMP）、通信协议。

WAVE 处理 WSMP，它是为 WAVE 环境中的优化操作设计的。WSMP 支持应用直接控制在传输消息中使用的物理层特征，如信道号和发送器功率。

一项应用也提供目的设备的 MAC 地址，可能包括一个广播地址。在提供商服务标识符（PSID）的基础上将 WSM 交付到一个目的地的合适（正确）应用。设计 WSM，是为了消耗最少的信道容量，并在 CCH 和服务信道（SCH）上得到支持。

4. WAVE 信道类型

WAVE 在两类无线电信道之间做出区分：单 CCH 和多 SCH。

缺省情况下，WAVE 设备工作在 CCH 上，这是为短小的、高优先级应用和系统控制消息保留的。除了这些流量类型外，在 CCH 上发送系统管理帧，见 IEEE 1609.4[5] 中的描述，而 IP 流量仅在 SCH 上得到支持。

为支持通用应用数据传递，在设备之间安排了 SCH 访问。

5. WAVE 管理实体和优先级

在管理平面中定义了 WAVE 管理实体（WME）。WME 使用应用优先级选择要服务哪个应用。以多种方式使用优先级概念。各应用有一个应用优先级等级，由 WAVE 联网服务使用，帮助确定哪些应用首先访问通信服务。一次冲突的一个例子将涉及两个应用，每个应用都有同时需要在一条不同信道上宣告或加入一个 WBSS。

此外，低层使用一个独立的 MAC 传输优先级，对在媒介上传输的分组划分优先等级。

IP 分组被指派与产生应用的流量类相关联的 MAC 优先级。WSM 分组的 MAC 优先级是在逐分组基础上由产生应用指派的[4]。

6.6 无线车载网络中的挑战

车载通信是极具挑战的。这是由几项因素导致的，其中有一个快速变化的环境、频谱状况中的动态改变（由于高移动性导致的）、干扰（由于邻居网络设备导致的）、多模态通信以及无线电需求中的改变（由多样化车载应用产生）。

在这个语境中，应该解决不同的研究挑战问题，以便提供可靠的数据传输和低延迟无线通信。

这些挑战涉及但不限于：

1）用于后来检索的数据的持久和可靠存储：复杂查询处理和联网协议，可高效地定位和检索关注的数据（如在某个时间和位置，找到一次碰撞周围的所有车辆）。

2）位置感知：从车辆采集的数据和车辆消耗的数据是高度位置相关的。这个性质对数据管理和安全组件的设计具有直接的隐含意义。数据缓存和索引应该聚焦在将位置作为第一位的性质。为维护隐私和防止篡改，数据传播必须是位置感知的。

3）海量分布的数据库：对数据库的创建和维护，应该定向投入特别的工作，这些数据库应该临时地存储可共享的内容。

4）拥塞控制：在绝大部分城区，交通拥塞是一个常见问题。考虑到最小化拥

塞的重要性，对监测高速路速度和交通流，应该投入大量关注，尝试解除交流拥堵和碰撞危险。在这点理解的基础上，在无线车载网络中应该实施预防性的拥塞控制。主要目的是规范和约束车辆交通。

5）提供服务质量：安全应用具有关键时延需求。由安全应用支持的时间敏感数据，必须在一个给定时间窗口内能够检索或传播到期望的位置。因此，为优先处理这些应用，应该施用特殊的机制和适合的呼叫接纳控制。

6）垂直移动能力：在重叠的异构接入网络语境中，必须设计特定的战略，控制可用接入网络之间垂直切换的触发，这会影响运行在移动用户的设备上应用会话的整体性能。

7）到数据消费者的高效路由：由于高度移动环境和高度分隔的网络，特别当智能车辆渗透率较低时，VANET 是由短路由寿命所刻画的。因此，应该设计特殊的路由协议，传播紧急的安全数据。

6.7 小结

最近观察到的无线数据通信技术的快速演化，为在支持车辆安全应用而利用这些技术方面，创造了丰富的机遇。

为避免碰撞及提高未来无线网络的容量和覆盖的目的，在支持车辆间通信方面，VANET 将扮演一个重要角色，其中采用如下做法：

1）在系统变得过载的热点处，补偿支援现有的蜂窝基础设施。

2）扩展蜂窝基础设施的覆盖区域，做法是支持一个范围外的车辆通过多跳转发它的数据（直到一个 BS 是可达的）。

3）以高于基于基础设施网络的速率在车辆间交换紧急数据。

另外，车辆和路侧基础设施之间的无线通信代表车辆通信的一种重要类型。基本思路是拓宽司机对视野外的感知范围，并进一步地以自治的辅助应用对司机提供帮助。

事实上，各 RSU 具有网络的一个全局视图，提供有用的信息，并有助于传播关键数据。这些智能的基于基础设施的单元，可就未来的危险道路状况、障碍、行驶路线不定的司机等，提醒司机。

在本章中探索了 V2V 和 V2I 通信的性质，讲解并讨论了应用的种类也说明了无线车载网络的架构。

在对聚焦车载网络的实际研究工作尝试做一个概述的过程中，给出了 V2V 和 V2I 通信所展望的标准。更精确地说，描述了 CALM、C2C-CC 和 WAVE 标准。最后，将焦点放在与无线车载网络有关的研究工作和挑战。

参考文献

1. CAR 2 CAR Communication Consortium, "Overview of the C2C-CC system," 2007.
2. Committee SCC32 of the IEEE Intelligent Transportation Systems Council, "IEEE 1609.1 draft standard for wireless access in vehicular environments (WAVE)—WAVE resource manager," 2006.
3. Committee SCC32 of the IEEE Intelligent Transportation Systems Council, "IEEE 1609.2 draft standard for wireless access in vehicular environments (WAVE)—security services for applications and management messages," 2006.
4. Committee SCC32 of the IEEE Intelligent Transportation Systems Council, "IEEE 1609.3 draft standard for wireless access in vehicular environments (WAVE)—networking services," 2006.
5. Committee SCC32 of the IEEE Intelligent Transportation Systems Council, "IEEE 1609.4 draft standard for wireless access in vehicular environments (WAVE)—multi-channel operation," 2006.
6. C. Chia-Hsiang, C. Chih-Hsun, L. Cheng-Jung, and L. Ming-Da, "A WAVE/DSRC-based intersection collision warning system," *Proceedings of the IEEE Ultra Modern Telecommunications & Workshops Conference*, 1–6, 2009, doi: 10.1109/ICUMT.2009.5345520.
7. Ching-Yao Chan, "An investigation of traffic characteristics and their effects on driver behaviors in intersection crossing-path maneuvers," *Proceedings of the IEEE Intelligent Vehicles Symposium*, 781–786, 2007, doi: 10.1109/IVS.2007.4290211.
8. A. Dogan, G. Korkmaz, Y. Liu, F. Ozguner, U. Ozguner, K. Redmill, O. Takeshita, and K. Tokuda, "Evaluation of intersection collision warning system using an inter-vehicle communication simulator," *Proceedings of the 7th IEEE Intelligent Transportation Systems Conference*, 1103–1108, 2004, doi: 10.1109/ITSC.2004.1399061.
9. O. Dousse, P. Thiran, and M. Hasler, "Connectivity in ad-hoc and hybrid networks," *Proceedings of the IEEE Infocom Conference*, **2**, 1079–1088, 2002.
10. A. S. Farahmand and L. Mili, "Cooperative decentralized intersection collision avoidance using extended Kalman filtering," *Proceedings of the IEEE Intelligent Vehicles Symposium*, 977–982, 2009, doi: 10.1109/IVS.2009.5164413.
11. M. Fornasa, N. Zingirian, M. Maresca, and P. Baglietto, "VISIONS: a service oriented architecture for remote vehicle inspection," *Proceedings of the Intelligent Transportation Systems Conference*, 163–168, 2006, doi: 10.1109/ITSC.2006.1706736.

12. S. Giri and R. Wall, "A safety critical network for distributed smart traffic signals," *IEEE Instrumentation & Measurement Magazine*, **11(6)**, 10–16, 2008, doi: 10.1109/MIM.2008.4694152.

13. W. Hao, "Analysis and design of vehicular networks," PhD thesis, Georgia Institute of Technology, 2005.

14. S. Inoue, K. Shozaki, and Y. Kakuda, "An automobile control method for alleviation of traffic congestions using inter-vehicle ad hoc communication in lattice-like roads," *Proceedings of the IEEE Globecom Conference*, 1–6, 2007, doi: 0.1109/GLOCOMW.2007.4437828.

15. Institute of Electrical and Electronics Engineers, "IEEE draft amendment to standard for information technology—telecommunications and information exchange between systems—LAN/MAN specific requirements—part 11: wireless LAN medium access control (MAC) and physical layer (PHY) specifications: amendment 3: wireless access in vehicular environments (WAVE)," 2007.

16. "ISO TC204 WG16 portal of ESF GmbH," www.tc204wg16.de.

17. J. Jakubiak and Y. Koucheryavy, "State of the art and research challenges for VANETs," *Proceedings of the 5th IEEE Consumer Communications and Networking Conference*, 912–916, 2008.

18. B. Mohandas, R. Liscano, and O. Yang, "Vehicle traffic congestion management in vehicular ad-hoc networks," *Proceedings of the IEEE LCN Workshop on User Mobility and Vehicular Networks*, 655–660, 2009, doi:10.1109/LCN.2009.5355052.

19. Official web page of the ISO TC 204 working group 16, www.CALM.hu.

20. C. Olaverri-Monreal, P. Gomes, R. Fernandes, F. Vieira, and M. Ferreira, "The see-through system: a VANET-enabled assistant for overtaking maneuvers," *Proceedings of the IEEE Intelligent Vehicles Symposium*, 123–128, 2010, doi: 10.1109/IVS.2010.5548020.

21. W. Pattaraatikom, P. Pongpaibool, and S. Thajchayapong, "Estimating road traffic congestion using vehicle velocity," *Proceedings of the IEEE ITS Telecommunications Conference*, 1001–1004, 2006, doi: 10.1109/ITST.2006.288722.

22. S. Thajchayapong, W. Pattara-atikom, N. Chadil, and C. Mitrpant, "Enhanced detection of road traffic congestion areas using cell dwell times," *Proceedings of the IEEE Intelligent Transportation Systems Conference*, 1084–1089, 2006, doi:10.1109/ITSC.2006.1707366.

23. J. Tian and K. Rothermel, "Building large peer-to-peer systems in highly mobile ad hoc networks: new challenges?" *Technical Report 2002*, University of Stuttgart, 2002.

24. P. Varaiya, "Smart cars on smart roads: problems of control," *Proceedings of IEEE Transactions on Automatic Control*, **38(2)**, 195–207, 1993, doi: 10.1109/9.250509.

25. Z. Wang, L. Kulik, and K. Ramamohanarao, "Proactive traffic merging strategies for sensor-enabled cars," *Proceedings of the ACM International Workshop on Vehicular Ad Hoc Networks*, 39–48, ISBN: 978-1-59593-739-1, 2007.
26. H. Wu, R. Fujimoto, and G. Riley, "Analytical models for information propagation in vehicle-to-vehicle networks," *Proceedings of the 60th IEEE VTC Conference*, **6**, 4548–4552, 2004.
27. H. Wu, J. Lee, M. Hunter, R. Fujimoto, R. Guensler, and J. Ko, "Simulated vehicle-to-vehicle message propagation efficiency on Atlanta's I-75 corridor," *Transportation Research Record (TRR)*, 2005.
28. Z. Yunpeng, L. Stibor, H. J. Reumerman, and C. Hiu, "Wireless local danger warning using inter-vehicle communications in highway scenarios," *Proceedings of the 14th European Wireless Conference*, 1–7, 2008, doi: 10.1109/EW.2008.4623905.

第7章　下一代 IPv6 网络安全：步向自治的和智能的网络

Artur M. Arsénio，Diogo Teixeira，João Redol

就下一代 IPv6 网络而言，将必须解决重要的安全挑战。IPv6 带来了网络安全的一些问题，原因是它将不再需要网络地址转换机制。但也将出现其他问题，也许具有更大的影响，涉及多媒体内容（交互式的和个性化的）的安全传输，经常通过对等网络传输。事实上，对等流量总是占有整体因特网流量的一个巨大份额，IPv6 网络的未来解决方案将需要管理所有可用资源，以便依据用户的通信概要，使用公平的规则，对用户实施缴费。因此得到有关因特网流量行为的信息，对于管理、监测和运营活动（如对客户使用的应用和协议的识别），是基础性的。但是，这种识别的主要障碍是网络设备的能力缺乏扩展性。特别地，它们需要针对这个目的而分析所有的网络分组。这项任务是极度紧迫的，在大型网络和高速下几乎是不可能的，原因是它们有数百或数千名客户。此外，采用 IPv6，期望这样的网络甚至变得更大，原因是在“物联网”上，所有设备（传感器、仪表等）都将公开地连接到因特网。如此，为克服这个重大的规模问题，人们已经提出了流量采样的战略做法。本章给出用于用户剖析和安全目的而监测流量领域中的不同工作。本章也给出下一代 IPv6 网络的一种解决方案，这种方案使用选择性的过滤技术与引擎流量深度分组检测（DPI）相结合的做法，以此识别客户最频繁使用的应用和协议。由此，使因特网服务提供商（ISP）以一种可扩展的和智能的方式优化它们的网络是可能的。

7.1　引言

7.1.1　背景

自因特网诞生、实现和被采用以来，因特网协议（IP）网络（如因特网）的快速扩散和增长与连接到网络的设备数量的爆炸式增长和全球蓬勃发展相耦合，使 IP 网络用于多种目的。与这种增长平行发展的因特网服务提供商（ISP），在多个层面上面临许多问题，特别与安全担忧相关，还面临 IP 网络被少数用户不公平使用的问题。如今，IP 网络几乎被用于所有领域，从因特网到公司到私有和个人层面，连接可想象到的所有类型的设备：计算机、打印机、智能电话、游戏控制台、电视、传感器等。但这种连通性仍然受到缺乏 IP 地址将设备连接在一起的阻碍。

事实上，因特网协议版本 4(IPv4) 地址不足以为所有设备分配一个公开的地址。新的因特网协议版本 6(IPv6) 已经正在解决这个问题，当前正在部署，特别在网络核心和城域网中得以部署。

与前述的增长相关联的是，在网络复杂性方面也出现了增长。在这个语境中，随着配置和维护达到非常高的和令人窒息的复杂水平[1]，在大型规模上理解这些异构网络的动态性，正变得日渐困难。IPv6 目标也在于解决这些问题，如路由性能改进。

1. 流量拥塞

与这种快速增长关联的一个主要的且是人们所不期望的因素（损伤了网络的性能）是流量拥塞，这是分组交换网络（如 IP 网络）的一项主要担忧。无论何时，当每个时间单位进入一个网络部分的流量总量大于网络单元处理和/或转发这种流量的容量时，就发生流量拥塞，这会导致路由器处的缓冲满和接下来的分组丢失。网络拥塞归结为网络上带宽可用性的问题，随着时间推移，这个问题会恶化，不管为确保服务质量而开发各种机制、技术和算法为何，都会恶化。过载和网络拥塞是由服务和应用的丰富性与传统使用模型的隐含意义所导致的，如各种对等(P2P) 应用和技术[2]，以及视频共享（如优酷）、生活广播站点和实况视频流化(如 Ustream、JustinTV、LiveStream)、在线游戏的快速增长［如魔兽世界（World of Warcraft)］等的多媒体站点的日渐增加的常见使用。

因特网协议上的话音（VoIP）如今也是因特网上的一项主导服务，就流量时延和抖动方面提出带约束的需求。而且对融合的网络服务，存在日渐增长的比较强烈的消费者呼声。

随着在未来连接到因特网的 IPv6 设备（如传感器）数量的巨大增长，人们预计在流量方面会有进一步的巨大增加，这导致网络拥塞。新的交互式和个性化多媒体内容的出现，将要求不太会是广播的而是更多单播的连接（从终端到内容提供商），这进一步非常显著地增加了网络流量。此外，多媒体传输方面使用 P2P 技术，将对电信运营商实施的网络控制产生偏见。所以，网络拥塞将是一个重大问题。拥塞经常是由于在一个或多个网络节点中不能将流量数据流调节到相对于服务速率而超出分组到达率，才产生的。这种超出导致网络节点处流量的不平衡，其中一组资源过载，而另一组资源可能是欠利用的[3]。

在过去的主导性思路是，拥塞可采用信道中传输速度的大量增加、通信节点处理能力的增加以及为存储分组而使用大型缓冲等措施得以简单的解决。但是，参考文献［4］的作者表明，这些规程单独都不构成一项高效的解决方案。

当前，在因特网上发生的拥塞，主要是由于通过它的流量的不可预测的和混乱的本质所导致的。基本上而言，当前的拥塞控制机制可被分类为两组：第一组通过相同资源的动态重新配置而增加资源的可用性，而最常使用的第二组，就资源的可用性方面，降低了需求[5]。

2. 网络安全

因特网是自然的一个持续变化的环境，这使它的客户依据其需要和期望而创建和调整它们的技术。在某些情形中，这种突变变得有点烦人，原因是随着因特网的出现，犯罪活动也变得相关起来了。在线犯罪，以及结果为计算机安全和联网问题，是当前ISP的另一项主要担忧。

在因特网上正在实施和实践非法和不正当的活动，如有版权材料的分发、交换和共享，免费地流化“支付TV频道”。

拒绝服务（DoS）攻击正在增加，为运营商造成颠覆性的结果和成本高昂的网络修复和维护。事实上，当前各公司报告非常大数量的非授权网络访问。事实上，多数防火墙每年要求关键性的补丁。

考虑到ISP方面的这些主要担忧，则至关重要的是知道每名用户在做什么，即用户使用的技术和应用。同样重要的是确定用户是否在利用网络实施犯罪活动。由此，变得清晰的是，必须产生这样的算法，来分析用户所产生的网络流量，并创建用户概要，从而使各ISP可依据所产生的概要而实现安全和商务策略。这种工作落在符合政策的网络管理领域内。

3. 自治和智能网络的安全动机

在过去数年间，IP网络（如因特网）的增长（与这些网络上用户的爆炸式增长相耦合）以及新应用和服务的出现，使网络流量在量和多样性方面出现增长。网络变得更加复杂，由此流量刻画和监测正逐渐成为流量工程的重要工具，这使网络运营商具有有关网络使用的多样化信息。

在对通信网络的一种更动态和高效架构需要的基础上，以及预期未来问题重要性的基础上，对开发智能机制存在一项需要，这些机制可产生一些输出，实际上可辅助ISP早期决策。

这些系统必须能够以一种可扩展的方式监测一个大型IP网络，并预防性能的降低，后者可能降低服务质量。同样令人期望的是，它支持判定一个网络客户是否正在将网络用于不正当的活动和犯罪目的。

这样的安全和性能需求（与互操作性一起）是IPv6广泛采用的一个前提条件。从IPv4网络过渡到IPv6，就如下两方面影响提出额外的担忧：

1）在IPv4网络安全上支持IPv6。

2）使设备和网络工作在双栈或隧道模式。

因此，必要的是高效地“预测”流量的变化及其对下一代IPv6网络服务质量和安全的隐含意义，这要考虑到由IPv6与IPv4网络互操作带来的两项前述担忧。当将最近发展（监测）的一项统计分析与产生未来状况的预测能力（预测）组合使用时，在本章中将进一步分析它们的影响。

7.1.2　下一代 IPv6 网络

从 IPv4 切换到 IPv6，正在对各组织施加新的挑战，这是就他们的网络实施防御的方式而言的。一些厂商防御这种情况，将导致从封闭的网络（默认地，对进入网络流量的怀疑是根本）步向开放的网络。

另外，正在成熟的 IP 网络，面临着纯 IPv4 流量、纯 IPv6 流量以及混合流量和隧道式 v4/v6 组合流量的复杂性。必须处理多媒体数据，包括数据、话音和视频。诸如 DoS 的网络攻击是持续存在的线索。存在处理遗留问题的需要，诸如 IPv4 网络地址转换（NAT），所有这些都必须由一项可行的 IPv6 网络安全策略加以解决。

1. IPv6 网络安全威胁

IPv6 是从 IPv4 演进而来的，但对万维网安全几乎没有带来什么改进，原因是后者关注在传输控制协议（TCP）/IP 模型的相当高的层处的应用安全，这个层比 IPv6 网络层要高，处于应用层。因此，就局域网（LAN）攻击：地址解析协议（ARP）与邻居发现协议（NDP）攻击、动态主机配置协议（DHCP）与 DHCPv6 分片攻击以及 DoS 攻击等而言，在 IPv4 和 IPv6 安全之间存在几项相似性。IPv6 首部的结构［与 IPv6 对因特网控制消息协议版本 6（ICMPv6）的强依赖一起］产生了另一个弱点。在 IPv6（相比在 IPv4）中，比较容易过滤未分配的地址，原因是大型 IPv4 地址空间分片导致 IPv4 不容易处理这个问题。就 IP 安全（IPsec）如何比较容易地实现方面，IPv6 提供一些优势，原因是 IPv6 不使用 NAT。移动 IPv6 为保障移动通信的安全提供新的机制。从 IPv4 到 IPv6 的过渡实现也给出一些新的弱点，这些可被攻击所利用。

相对于 IPv4 而言，这些新产生的差异不对 IPv6 造成重大弱点。在 IPv6 上目前识别出的主要安全弱点有[6]：

1）IPv6 首部的处理（扩展/选项首部）。路由首部类型 0，这是类似于 IPv4 上“源路由”的一个概念，支持旁路防火墙（用来保护网络的硬件或软件组件），支持通过替代路径中继流量。

2）当使用 IPv6 到 IPv4 隧道时，通过使用 IPv6 到 IPv4 网关旁路 IPv4 ACL，此时可旁路访问控制列表（ACL）。因此，计算机可创建到 IPv6 因特网的隧道，这可旁路所有当前的纯 IPv6 安全防护措施。

3）入侵检测系统（IDS）/入侵防御系统（IPS）对大型扫描搜索空间的影响。IPv6 子网使用大型的地址空间（2^{64}个地址）。这会导致针对 IDS 的一项不可完成的任务，原因是 IDS 必须在这样一个大型空间上搜索 TCP 和用户数据报协议（UDP）流量。正是模块化的 IPv6 首部结构导致了 IDS/IPS 的额外问题，原因是设计攻击签名变得更加困难。付出额外的计算成本，标准强制 IPv6 网元（NE）实现 IPsec。但 NE 可选择是否使用 IPsec，解密所有 IPsec 流量并不总是可能的。另外，忽略一些加密流量也许导致错失检测攻击，这在一些安全策略中也许是不可行的。但正如

将在本章中看到的，基于自适应流量采样的一种解决方案，用于在网络周边或内部有选择性地过滤分析用的样本，可有助于防御这些类型的攻击。在一些场景中，这可能是最佳的解决方案，特别当目标的焦点是用户剖析而不是直接的强安全措施时更是这样的。

4）NDP 毒化（不是 IPv4 ARP 毒化）。在 TCP/IP 栈的层 2 上，在 IPv4 安全内存在弱点，即 ARP 上的弱点，可被用来毒化以恶意地重定向流量。同样的问题在 IPv6 NDP 中也存在。

5）对 DoS 攻击的 IPv6 移动性弱点。

同时，因特网工程任务组（IETF）是因特网标准化组织，标准化了 IPv6 的一些细小演化修改，解决前述的一些问题：

1）安全邻居发现（SEND）协议的建议。这个安全协议采用密码学方法，保障将一个 IPv6 地址映射到一个以太网媒介访问控制（MAC）地址的动态发现的安全。

2）路由首部类型 0 的废弃使用，这避免一些 DoS 攻击的可能性，并禁止 ACL 和防火墙的旁路做法。

2. IPv6 和 IPv4

（1）用于所有事物的地址

在有关过渡到下一代 IPv6 方面，一些组织总是存在一些滞后，但就剩余空闲 IPv4 地址的结束（时间）将总是不可避免的。随着用于环境和工业检测的大型传感器网络、用于个人保健监测的个域传感器网络以及用于无缝产品跟踪的新的物流系统等分布式应用的部署，128 比特的大型 IPv6 地址空间（而不是 IPv4 的 32 比特地址），对于连接每台网络服务器、笔记本电脑或台式计算机、智能手机、智能家庭、三重播放解决方案、万维网摄像机以及连接到因特网的任何其他设备，都是至关重要的。当然，因为防火墙规则集和 ACL 现在必须使用这些大尺寸的 IPv6 地址，所以它也会影响性能。

（2）IPv4 网络安全和 NAT

就 IPv4 网络安全而言，一个组织典型地建立一个周边防御，将机器放在它的 LAN 上，多数或所有机器都被指派本地 IP 地址。防火墙被用来控制去往网络/来自网络的进入和/或外发流量［经常的情况是，公开可访问的机器被放置在一个隔离区（DMZ），使用两个防火墙或单个三臂防火墙］。流量检查被用来检测网络使用的异常模式，或具有类似于那些已知攻击之模式的流量。网络终端经常装备有抗病毒和抗垃圾软件。

安全通信协议，如传输层安全/安全套接层（TLS/SSL），被广泛地用在商务应用上。各组织也利用虚拟专网（VPN）进行内部网资源的远程访问。也存在标准化的网络访问控制机制，如 IEEE 802. 1x。

一个组织典型地有有限数量的公开 IP 地址（由一个 ISP 提供），并使用 NAT

向组织中的所有机器赋予因特网访问。也可能存在处理层5、层6和层7攻击的一种IPS。

NAT不仅为该组织之LAN的所有设备提供丰富的个体专用IP地址，而且它也为这些设备提供安全，原因是防火墙外的任何人不能访问LAN上的各设备。这也对P2P流量提出一些问题，原因是对等端需要额外的组件（如一个STUNT注册服务器）从外部发起通信。

所以，因为防火墙内部的每台设备都对外部的潜在攻击者是隐藏的，所以这使攻击非常困难。

NAT使攻击者更难以检查网络（为了推断设备的活动）。但NAT对许多类型的攻击（如钓鱼）不提供保护，这些攻击是从组织的防火墙内发起的，不受怀疑地将恶意软件下载到用户的机器上。从内部扫描网络，对一名攻击者而言是相当简单的，他在恶意软件被攻破的机器上采用标准的检查攻击，像Nmap即可。

简言之，NAT阻塞源自外部的连接尝试，但采用一个有状态的防火墙，可取得一个类似的目标；旁路NAT并到达有一个私有IP地址的一台机器，存在各种方式，例如使用反向隧道或一个外部代理（这是Skype采取的方式）。

NAT隐藏网络拓扑，但确实存在可帮助一名攻击者采集有关网络设置信息的技术。例如：

1）统计ID字段或TCP时间戳。

2）分析存活时间（TTL）或电子邮件请求评述（RFC）882首部。

采用IPv6（因此没有NAT），存在一个网络上任意设备和任意外部设备之间端到端的可能优势，原因是每台设备都有其自己的独特IP地址。这个模型非常适合P2P应用（也产生新的担忧）的快速发展和商务机会，原因是运营商可利用这样的流量，对其进行合适的收费，而不会遇到传输P2P流量（在没有为之付费的情况下）的劣势。

在必要的情况下，通过使其拓扑对攻击者可见，IPv6网络没有变得不太安全。由于在IPv6子网上存在大量地址（默认情况下，有2^{64}个地址），一名攻击者要完成一次网络扫描，将花费不可行的时间。这意味着，攻击者必须使用其他机制，像被攻破机器上的域名服务（DNS）记录分析或日志检查或netstat数据，以便抽取有用的信息，这使他们可攻击其他机器。

（3）后向兼容性

虽然在IPv4和IPv6之间就安全性不存在真正的区别，但IPv6并不后向兼容于现有的IPv4产品。因此，为保障IPv4和IPv6网络的安全，应该使用相同的工具，如防火墙、IDS/IPS、网络管理测量、为检测异常行为的行为分析以及许多应用安全产品，像防垃圾和防病毒，可检查网络消息。7.2节将进一步回顾这些安全工具。

但是，对IPv4和IPv6的同时支持产生一些安全担忧。虽然运行一个双栈环

境，其本身并不导致任何安全问题，但是网络对这两种协议的安全问题变得脆弱起来。也存在这样的攻击，它们利用一种协议攻击另一种协议。

3. IPv6自动配置和信任

IPv6为IP网络带来自动配置，因此需要解决自动配置安全挑战。这样的一个挑战是信任（例如，谁透明地配置网络和设备）。另一项挑战涉及动态网络配置，所以需要部署新的智能和自治算法。

所以，对于下一代IPv6网络，与使所有外部流量不能进入网络的做法不同，一旦使用如下技术对流量进行过检查，则应该接纳这样的流量。这些技术如：

1）IP地址声誉解决方案，它监测流量，并阻塞低声誉得分的IP地址流量。

2）带有动态签名更新的一个IPS，其性能高度依赖于流量概要和配置。

7.1.3 本章结构

本章也将报告在卓越电信研究所（Instituto Superior Técnico）和诺基亚西门子网络公司（NSN）之间协作项目中所做的工作。NSN聚焦在几个领域，包括计算机联网以及自治和智能配置领域。这项工作研究流量分析的各种技术以及流量分析算法的设计。这样的算法输出用户概要的构造，之后与用户关联，所以ISP可通过概要或指派到客户的概要而将策略（如流量整形[7]）实施到客户，同时对对应于那些概要的签名外的流量（如P2P流量）实施额外缴费。

因此本章从结构上分成四节。本节介绍IPv6网络面临的主要安全问题和解决这些问题的动机。7.2节给出最新技术，方法是给出对IPv6网络合适的当前方法论的一项分析，这些方法当前也用在IPv4网络上。7.3节给出满足要求的下一代IPv6智能网络的一种架构，并给出构成该架构的不同组件的详细描述。最后，7.4节给出下一代IPv6网络安全的讨论和一些结论。

7.2 相关工作、工具和协议

本节给出涉及客户剖析和网络安全工具的标准及相关研究的概述，并给出与这个主题相关的商务解决方案。

7.2.1 入侵检测/防御系统概述

在数年间，计算机网络在规模上出现了巨大增长，被用在多个领域，如军事、金融和大规模电子商务等。由于用户和关键应用的巨大增长，以及新的网络攻击技术的出现，计算机网络为管理人员带来新的挑战。网络需要稳定性，原因是它们负责从多个源传输大量信息和数据（其中多数是机密）。由此，常见的情况是尝试攻击网络，目标是中断数据和信息的机密性、完整性和/或可用性。因此，在所开发的几项技术中，出现了IDS和IPS。从传统角度看，当攻击发生时，一个IDS识别

这些攻击。另外，一个 IPS 可先验地阻塞这种攻击并最小化损害。重要的是研究 IDS 和 IPS 的操作和架构，以解除其功能、算法、机制、优势和劣势的神秘面纱，并分析其应用于下一代 IPv6 网络的潜力。

一个 IDS 是一种安全管理工具，该工具可辅助并自动化监测网络事件的处理过程[8]。在参考文献［9］中，IPS 被定义为这样的设备（硬件或软件），这些设备具有检测攻击（已知的或未知的）并防御它们成功攻击的能力。术语 IPS 已经在文献中使用，作为 IDS 的演进，将防御攻击的功能添加到简单检测入侵者的功能上。因此，一个 IPS 也许本地作用在一次尝试的入侵上，防止它取得入侵的目标，同时最小化损害。

当前有两种不同观点。一种观点论辩，将响应功能添加到 IDS，并不能证明创造一个新的术语的合理性，而另一种观点则认为在系统中实现先验式阻塞措施，足以将其以另一种方式分类。为解决这些争议，使用参考文献［10］的术语，该文献表明，存在三类工具：IDS、具有主动响应的 IDS 和 IPS。在这个分类内，IDS 监测主机，目的是通过异常或签名检测可疑的活动，并在不干扰网络流量的条件下产生告警。带有主动响应的 IDS，目标是采取行动，间接地自动地关闭所检测到的可疑活动。一个 IDS 单独不能停止攻击，它需要其他机制的帮助才能实施这项功能。另外，IPS 具有与一个 IDS 相同的检测机制，但它可实时地和自动地停止一项可疑的活动，有无其他设备的帮助均可做到。

1. 检测方法

IDS 和 IPS 都是用三种检测方法产生告警或阻塞任何可疑的流量。检测方法可基于异常或签名或两者的一个混合体。

（1）基于签名的检测

这项技术基于使用一个数据库，存储某些攻击的模式，这些模式被用来与正在发生的可能攻击进行比较[11]。依据参考文献［12］，一个网络 IDS 签名是要在网络流量上进行搜索的一种模式。当一次攻击的一个签名对应于被观测的流量时，产生一次告警，否则记录一个事件。一次攻击的签名是在包含该攻击的分组特征基础上构造的。其中包括源/目的端口、序列号、协议标志（如 syn、ack），以及特别是一小片应用层（数据）[13]。

这个系统的劣势，一方面，是其侵略性本质（信息隐私问题），另一方面，则是仅能检测已知的签名和可能的变形。由此，总是保持一个最新的签名数据库，是非常重要的。

（2）基于异常的检测

基于异常的检测，特别适合于安全目的，它假定每个用户具有资源使用率的一个概要，目标是检测与这些模式的偏离，以便识别可能的攻击[11]。这种方法是识别在一台主机或一个局部网络上的不同行为（异常）。它假定各攻击不同于常规活动（合法的），并可由识别这些差异的系统加以检测。异常检测器构建概要，代表

用户、主机或网络连接[13]的正常行为。用来实现异常检测系统的理论范例有如下一些[11]：

1）检测阈值：该分析基于监测在一个给定时刻单一类型的活动，可能是一小时前无效登录次数或在一天中被删除的文件数。

2）基于概要的检测：该分析基于度量已知活动的一个不同类型集合。

3）统计方法：这由统计数据处理所表征，其中监测用户和过程活动，以便产生其概要。该系统周期性地产生某个概要异常水平的一个指示值。

4）神经网络：训练一个神经网络，将其应用到某些活动，产生一个概要。无论何时概要中发生一个给定的变化，则这就会为神经网络检测到，支持将这种状况识别为正常或异常（攻击）。这个系统的巨大优势是可能识别新的攻击技术，原因是该系统基于异常检测，神经网络能够将以前学到的知识做一般化处理。劣势是，为整个网络创建一个概要是太过复杂的。

(3) 混合检测

混合入侵检测（这种系统的一个架构例子如图 7.1 所示）是联合使用基于签名的监测方法和基于异常的检测方法。其目标是纠正这两种方法的每种方法所呈现的缺陷。

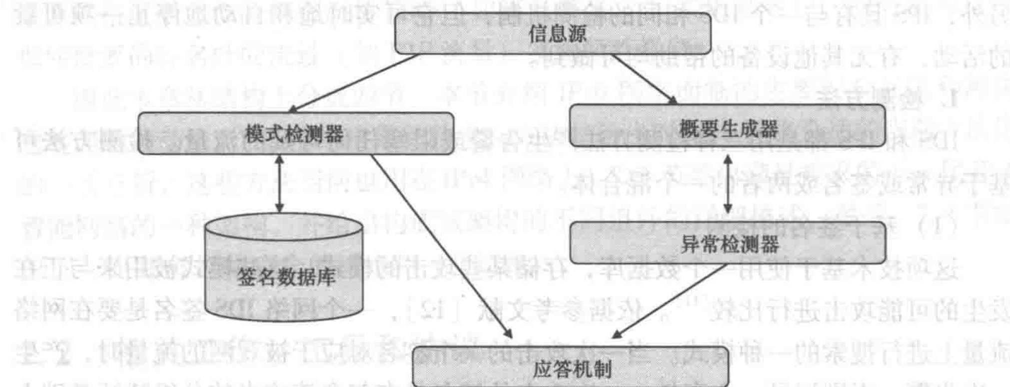


图 7.1　一个通用混合检测系统的架构[14]

2. 架构

IDS 或 IPS 的最著名架构有：

1）基于主机的架构：该系统被安装在一台主机上，但仅负责该主机的安全。这个架构包括一个主机入侵检测系统（HIDS）和一个主机入侵防御系统（HIPS）。HIDS 被用来检测和分析一台特定主机中的可疑活动。见参考文献［12］中所述，这种类型的检测涉及监测系统的网络活动、文件名系统、日志文件和用户活动。一个 HIPS 工作情况类似于一个 HIDS，并在攻击成功实施之前停止攻击。一个主动响应系统的这项巨大差异是，防御系统可直接访问应用和操作系统（OS）内核本身[10]。

2）基于网络的架构：一个或多个IDS或IPS传感器将负责各网络分段。这个架构包括一个网络入侵检测系统（NIDS）和一个网络入侵防御系统（NIPS）。一个NIDS目标是监测通过一个给定网络分段的所有流量，以便通过监测扫描、探测和攻击而识别威胁[12]。NIDS架构有一个或多个传感器，它们负责分析通过一个给定网络分段的流量。一个NIPS是一个线内设备，即它直接位于分组穿越网络时的分组路径之中[10]。通常的情况是，基于主机的IDS/IPS主要目标是内部用户，而基于网络的IDS/IPS主要将目标锁定在外部入侵者上，虽然存在例外情况。

3）分布式架构：这个架构可能有本地传感器和网络传感器。这个架构包括一个分布式入侵检测系统（DIDS）和一个分布式入侵防御系统（DIPS）。DIDS和DIPS具有本地传感器（HIDS）和/或网络传感器（NIDS）。这个架构主要区别于其他架构之处在于这样的事实，除了有各传感器要连接到的一台管理站外，这些传感器也可相互间交换信息。取决于需要，每个传感器的规则是可定制的[15]。

3. 采用ESP的IPsec

使用采用加密的IPsec，即在IPsec协议族上利用封装安全净荷（ESP），可使IDS和IPS工作更加困难。事实上，一个ESP提供诸如机密性、完整性和数据源发认证等服务，提供加扰TCP分组的能力。这可阻塞IDS/IPS针对监测目的而解密TCP分组。取决于所涉及的应用，这对IDS/IPS用途可能施加严重的限制。

7.2.2 监测网络流量

监测网络流量是网络主动或被动管理（可能有各种维度）的一项核心活动。可通过观测分组或流，实施监测。在科学共同体中进行了巨大努力，宏伟目标是深入理解各种应用的流量特征如何影响网络基础设施的行为。因此，测量战略对识别异常行为（如所产生的流量突然出现高的总量）是至关重要的。

现在给出网络管理和监测的一些关键概念。基于简单网络管理协议（SNMP）的网络监测，也基于NetFlow和sFlow协议，是专门为流量分析设计的。

1. 简单网络管理协议（SNMP）

SNMP是一种管理协议（TCP/IP应用层），其主要目的是管理一个网络的设备，并检测问题。在SNMP各RFC（特别是RFC 3411[17]和3416[18]）中可找到一个完整描述，这些RFC构成标准STD62[16]。

在基于SNMP的网络管理中，主要概念有管理器、代理和被管对象[19]。管理器是得到信息并控制被管对象的实体。这个角色可由单台主机担任，它允许一项管理应用。一个代理在被管对象上执行管理操作，也可将由这些对象发出的通知传输给管理器。运行在一个主机软件控制或交换机上的守护进程是代理的例子。一个被管对象是一个资源的表示，它受到管理。这项特征可以是一台网络设备甚至一条连接。被管对象是以属性或性质定义的，还有可被执行的操作、可被发送的通知以及它与其他对象的关系。在一个系统内的被管对象集，还有其属性，是管理信息库

(MIB)。每个 MIB 使用由1型抽象语法表示（ASN.1）定义的一个树架构，来组织所有可用的信息。由此，一棵树中的每片信息是一个带标签的节点。每个节点包含一个对象标识符（OID）和一个描述。一个节点可包含其他节点，如果它是一个叶节点，则它也包含一个值，被称作对象[19]。

SNMP 本身定义一组操作，用于一个代理和一个管理器之间的通信。每个代理维护一个 MIB，它反映由这个代理所管理资源的状态。所以一个管理器可监测这些特征（通过读取 MIB 对象的值）甚至控制这些特征（通过修改它们的值）[20]。一个代理也可发送包含信息的警报（在没有查询的情况下），其中采用操作 TRAP 和 INFORM，简单地称作陷阱。一个管理器也可向一个代理发送配置改变的请求，其中使用 SET 操作。在参考文献［16］中完整地描述这个协议。SNMP 是传统地基于一个客户端-服务器架构的，它使用 UDP。

2. NetFlow 和 sFlow

一个分组流（流量流、分组流或网络流）可被定义为一组分组，它们使用相同的协议、源地址和目的地址[21]。针对一个分组流，存在多种不同的定义。RFC 2722 将一个流定义为一个呼叫或连接的一个人工逻辑等价物。RFC 3697 将一个流定义为从一个特定源发送到一个目的地（单播、任意播或组播）的一个分组序列，其中源将之标记为一个数据流。RFC 3917 将一个流定义为在一个给定时间间隔期间通过网络上一个观察点的一个 IP 分组集。

网络设备制造商已经开发资源并在其产品包括这些资源，这支持获取网络流量的更详细信息。诸如 NetFlow[21] 和 sFlow[22] 的协议被用来得到有关分组流的信息。在这些协议的帮助下得到的信息，可使用诸如 flow-tools㊀[23] 和 ntop㊁等工具加以存储和分析。

NetFlow 是由思科开发的一个协议[21]。支持 NetFlow 的一台路由器实现能够统计分组流的一个代理。NetFlow 的一条流被定义为一个给定源和目的地址之间的一条单向分组流。更具体而言，一条流被看作带有相同字段的一个包集合：源 IP 地址、目的 IP 地址、源端口、目的端口、协议、服务类型和输入接口。除了用来定义流的字段外，该代理能够存储数据，诸如包的源自治系统（AS）和目的 AS。NetFlow 代理在路由器中为每条活跃的流维护一个缓存，为每条新的分组增加分组数和字节数。之后使用 NetFlow 协议，将这些表项输出到一个收集器。在思科路由器中使用的 NetFlow 代理也能够通过一个特殊的 MIB 使用 SNMP 提供缓存信息。因此 NetFlow 支持汇总并采集有关通过一台路由器之流量的统计信息。因为仅存储流数据而不是每个包的内容，所以可为大量流量存储信息。由所存储的数据，就可能得到如下信息，诸如目的地为一台服务器的一个特定端口的字节数和流数，或由一

㊀ flow-tools：http：//www.splintered.net/sw/flow-tools/(最近访问时间：2011年5月27日)。

㊁ ntop：http：//www.ntop.org（最近访问时间：2011年5月27日）。

台主机产生的总流量。

sFlow是用于高速网络流量监测的一项技术，这是由RFC 3176[22]定义的。在sFlow代理中，不必要维护流信息，所以代理部署得到极大简化。另外，将规范设计为支持在每秒吉比特量级或更高的速度上接口的精确监测[22,24,25]。sFlow，不同于思科的NetFlow，实施要被分析的分组的采样，这些分组穿过交换机或路由器，降低了监测分组所要求的处理（能力）。在监测节点，它运行代理，代理实现采样机制。在中央服务器处，sFlow收集器接收和存储数据报，数据报要被处理（见图7.2中的数据报例子），数据报是由sFlow代理发送的，以便进行进一步的分析。sFlow使用两种不同的采样方法：计数器采样或基于分组的采样。在计数器采样方法中，一个轮询间隔定义字节或分组计数器（穿越一个特定接口的字节或分组）的内容有多频繁地被发送到采集器。在基于分组的采样中，每N个分组捕获一次。这种形式的采样在结果方面不能提供100%精确度，但可以对应用必要的精确度产生一个结果。存在这样的图景，它使用分组的统计采样，以合理的准确度重构被采样的流量是可能的。被收集的采样数据通过UDP分组发送到采集器，由其IP地址和端口指定[22,24,25]。

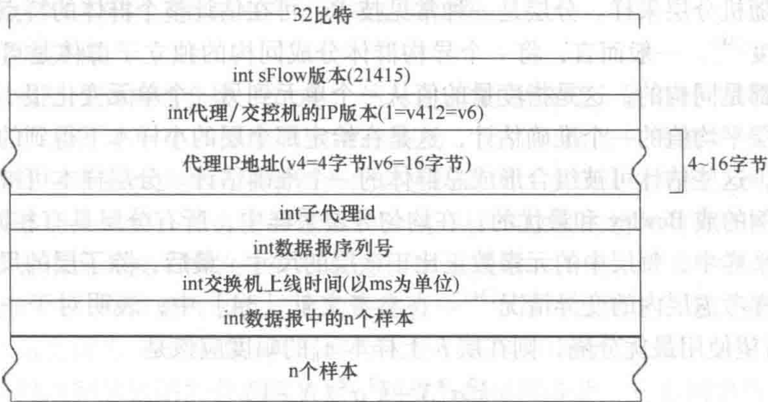


图7.2　sFlow数据报[22]

除了sFlow和NetFlow工具，也存在jFlow（也称作cflowd，由Juniper网络公司开发）和NetStream[26]（由华为技术公司开发）。jFlow是用于IP流量流的一种采样技术，并被看作非常类似于sFlow流的一项技术，当在一个接口上执行时，它支持在输入流中的分组被采样。NetStream是等价于NetFlow的一项技术，但用于华为的设备[26]。

7.2.3　分组采样和流采样

在路由器中现有工具的帮助下（如思科NetFlow），监测活动变得常见了，除了这样的事实外，仍然持续存在几个问题。基于测量的（分组或流）流量监测的当前主要障碍是相对于链路容量缺乏扩展性。换句话说，在具有非常高容量的链路上监测流量导致产生巨量数据[27]。随着链路容量和流数量的增加，为经过路由器

的每条流维护计数器，在计算上和经济上都变得代价高昂起来[28]。

因此，最近提出了几项采样战略，作为优化包选择（对于统计流）[16,27,29-31]或流选择（对于原始流量的统计分析）[32]的一种方式。简单的采样过程（均匀的）不能提供充分的结果，原因是对IP流的分组和字节而言，它们一般遵循Pareto分布，也称作长尾分布（长的尾部）[33]。一些现有的采样技术依赖于流尺寸，其中仅统计相对较大的流。

因此，就包的采样而言，本质上要严格地探索和分析各种现有的方法。

1. 分层采样

这里描述应用于流量分析的分层采样技术及其用于降低被采样数据量的情况。在分层采样[34]中，由N个单元组成的一个群体，首先被分成有N_1，N_2，…，N_L个子单元组成的子群体。子群体是不重叠的，总体覆盖整个群体，满足N_1，N_2，…，$N_L=N$。各子群体被称作层。为得到分层的全部优势，必须知道N_h个值。在确定层之后，在每层中应该选择一个样本，选择是独立地在不同层中做出的。在层内样本的数量分别被称作n_1，n_2，…，n_L。当在每层中选择简单的随机样本时，整个过程被称作随机分层采样。分层是一种常见技术，可在估计整个群体的特点中提供增加的准确度[34]。一般而言，将一个异构群体分成同构的独立子群体是可能的。如果所有层都是同构的，这是指度量的值从一个单元到另一个单元变化很小，则可得到任何一层平均值的一个准确估计，这是在给定那个层的小样本下得到的[34]。

最后，这些估计可被组合形成总群体的一个准确估计。分层样本可被分类为均匀的、比例的或Bowley和最优的。在均匀分层采样中，所有分层具有相同的尺寸，而在比例采样中，每层中的元素数正比于该层的尺寸。最后，除了层的尺寸外，最优分层采样考虑层内的变异情况[34]。在参考文献［34］中，表明对于一个给定的n，如果希望使用最优分隔，则在层h上样本n'_h的幅度应该是

$$n \geqslant \frac{k^2\sigma_1^2 N - k^2\sigma_\sigma^2(N-1)}{\varepsilon^2(N-1)+k^2\sigma_1^2}$$

式中

$$\sigma_\sigma^2=\frac{\sum N_h\sigma_1^2}{\sum N_h}-\left(\frac{\sum N_h\sigma_h}{\sum N_h}\right)^2;\ \sigma_1^2=\frac{\sum N_h\sigma_1^2}{\sum N_h}$$

k——正态分布的$(1-\alpha)$量；

ε——准确度误差。

在不同层间（由一个最小方差的条件得到）为样本元素的分布建立一个准则，则在n个元素的一个样本上，层h中元素n_h的数量是由如下表达式给出的[34]：

$$n_h=n\left(\frac{N_h\sigma_h}{\sum N_h\sigma_h}\right)$$

式中 n——样本尺寸；

N_h——单元总数；

σ_h——层内的标准方差。

目标是确定必须被抽取的样本尺寸 n，以便估计这个群体的某个特征（如流量流的平均尺寸或其平均长度）[34]。

依据参考文献 [35]，分层采样是混合技术的一个例子。分层采样背后的基本思路是，使用有关特点相关性的先验信息增加估计的准确性，这些信息采用某种其他技巧探索是容易得到的。该先验信息被用来实施主要群体各元素的一种智能分组。所以，可得到样本尺寸的一个较好估计，甚至更可能的是，在不降低估计准确度的条件下，降低样本尺寸。多篇文章解决这种采样模式。在参考文献 [36] 中，作者们探索这种方法，作为在流层次描述流行为的一个工具。在参考文献 [37，38] 中，它使用分族分析技术 [即 K 均值和分族及大型应用（CLARA）] 并将之用于流量流的分层采样。在参考文献 [37，38] 中给出的结果清晰地表明，算法 CLARA 和 K 均值适合于基于"流时长"度量的分层实现。参考文献 [38] 表明，在不增加样本尺寸的条件下，使用分层采样技术如何提高估计的准确度。就被采样包数的可能降低方面，该文通篇探讨了不同的分层战略。该文表明，如果依据分组的维度对之分层，则样本尺寸可被急剧降低。

2. 自适应采样

采样率直接或间接地变换（对应）估计过程的准确度。使用低采样率，一些网络行为（例如异常）是不能准确地被检测的。另外，高的采样率产生大量的数据，这些数据后来必须提交给采集器并由其进行处理。在高流量期间，网络设备不能处理所要求的采样率，并会丢弃过量的分组。样本数量的增加可能影响总流量，原因是大部分情况下被采样的分组是通过 UDP 发送的。由此，重要的是避免拥塞。因此，明显的是，在准确性和性能之间存在折中。不是如此明显的是，如何选择最佳采样率，这会成为一项挑战，甚至成为一项不可能完成的任务。

在不同时段穿越链路的分组数方面，网络流量呈现变化。有关网络行为令人印象深刻的是流量的突发，原因是网络流量可由一个重尾分布所刻画[39]。简短而言，在没有活动（或低负载）的情况下，样本应该是比较广泛空间分布的，以便降低所消耗的带宽和存储的信息量。无论何时存在大量网络活动时，采样都应该是比较频繁的，从而不会丢失有关网络状态和性能的重要信息（虽然这代表了带宽的额外消耗）。

可用来实施自适应采样的两种技术有[40]：

1）线性预测：使用过去的样本来估计未来测量。这项技术是以一组规则来分组的，这些规则定义在采样期间要实施的调整（量），依据的是对正确或不正确预测的反馈。根据 Jurga 和 Hulbój[39]，当与其他方法比较时，线性预测提供足够的准确度。

2）模糊逻辑：使用过去的采样，计算算法的各种参数，由此自动地调整采样间隔。

就自适应采样的机制而言，一个极大群组的解决方案将焦点放在实现定制的采样方法上。Paxson 等[41]描述两种自适应的采样方法，来管理在一台网络设备上的处理器使用情况。一种方法使用有关处理当前使用情况的信息，来调整采样率。另一种方法使用分组的到达时间（可被用来预估一次流量突发）和有关处理时间（处理一个样本所需时间）的知识。Chaudhuri、Motwani 和 Narasayya[42]建议使用最小平方估计和某个启发式规则集，来确定采样率。Duffield、Lund 和 Thorup[43]描述一种流采样方法，这使我们可控制期望的样本总量，并最小化估计的方差。所提出的智能采样方法调整采样处理，并组合使用一条流（采样流尺寸选择的）的似然率。这个过程将焦点转移到“巨大”流，它对流量总量具有一个严重影响。参考文献［16］中的专利，在解决方案中包括一种采样率调整机制。Choi、Park 和 Zhang[25]提出一种自适应采样方法，它调整采样率，从而在没有过采样的条件下，限制流总量估计的误差。该方法使我们能够控制估计的准确度，这是测量效用和开销之间的一次折中。

参考文献［44］的作者介绍了黏滞采样，这是基于所存储记录数调整采样率的一种方式。Chen 等[45]提供为每条流动态调整采样率的机制，目的是维持一个均匀的相对误差。有关采样率的预测和调整存在许多其他方法，在准确度和复杂性方面它们存在显著差异。最简单解决方案之一是原始预测，它假定下一间隔的分组数将等于当前时段的分组数。它是这样一种解决方案，即几乎不要求计算，但丢失了估计的准确性。已经讨论过的多数预测方法均可用在这种情形中。但是，从设备内部网络访问一些数据是不可能的（作为队列的状态、分组到达率、实时资源利用率等）。可被用来预测未来的主要信息[39]有以前的分组计数和以前分组总量。

7.2.4 深度分组检测

深度分组检测（DPI）涉及穿越网络的各分组的透彻分析，不仅检查首部［像由浅层分组检测（SPI）所做的那样］，而且检查它们的内容。但是，来自因特网的分组不仅仅是通过添加单个由净荷数据形成的首部。事实上，在多层架构的每层处，要向负载添加一个首部，且净荷首部包含一个高层。因此，一个较好的定义是基于 IP 首部和 IP 净荷之间的边界上的。

由此，DPI 的定义可被定义为任何网络设备（将一条通信信道的端点上的终端排除在外）的动作，使用在网络层上部一个层上的任意字段，这是相比 SPI 而言的，SPI 仅检查一条分组首部的一部分[45]。现代网络设备利用 DPI 实现复杂的服务，如入侵检测和防御、流量整形、负载平衡、防火墙、垃圾检测和病毒检测等。DPI 是在分组上实施匹配准则的一种强大机制[45]。

一份技术报告[39]表明，Snort 可被看作一个软件 DPI，但它不能处理高速流量。这主要是由于顺序型冯·诺依曼架构的限制，同时由用于匹配的正规表达式的不佳优化导致的。在 DPI 的情形中，经常必要的是将模式与每个字节的偏移进行匹配。

极可能的是，多个签名必须与分组净荷相比较。由此，该过程要求大量比较操作。顺序型处理不适合这种操作模式，由于这个原因，要采用定制的并行方法。DPI应该能够至少提供标准指标数和有关位置的信息。它也应该支持模式的分组。仅应该测试一组模式以便确定分组是否属于那个组。一个首部分类器与一个净荷匹配器一起使用可构造一个更加高效的和准确的DPI系统。

因为顺序型架构不足以实施DPI任务，所以一些研究人员将焦点放在开发现场可编程门阵列（FPGA）的并行实现上。通常，为高效的多模式匹配，使用如下三种算法之一：布鲁姆过滤器算法、Aho-Corasick算法和Boyer-Moore算法。布鲁姆过滤器算法及其扩展是最常见的算法。这个算法使用多个哈希函数，并可能产生假阳性（但从来不会是假阴性）结果。在参考文献［45］中，实现并测试了一个DPI中采样分组的六种不同方式（即不变量随机采样、不变量机械采样、随机时间采样、机械时间采样、随机采样速度模式和速度机械采样），检测高可用网络中P2P数据的数据流。

7.3 IPv6网络安全和用户剖析的智能

这里给出了对采集网络流量样本和处理这种信息方面重要的几个工具和协议，以便能够做出有关安全和服务使用一般情况的决策。在7.1节也看到，这种剖析对下一代IPv6网络是特别重要的，原因是不仅不再使用NAT，而且由于IPv4迁移到IPv6，出现了额外的安全需求（和新的IPv6需求）。此外，IPv6将支持甚至更多的下一代服务，如基于P2P流量或云计算或大型无线传感器网络的那些服务。因此，设计智能IPv6网络是极端重要的，以便保障安全性，并为服务使用确定合适的用户概要。

本节给出一个可扩展的架构，该架构通过与一个使用概要关联，采用被采样网络流量的分析，能够推断一个特定网络客户端（用户）的行为。为做到这一点，对一个自动系统建议为实时流量分析，使用智能和可扩展算法。为满足如下需求，在策略的基础上，该系统自动地配置网络：

1）实时流量分析和捕获。

2）为分组捕获确定最优的采样率。

3）流量的分段（组织），这些流量是由客户捕获的。

4）将一个客户映射到一个特定概要。

一个用户概要由用户最频繁使用的服务/技术集（即P2P站点、视频共享、在线游戏站点等）组成。这种服务的使用产生流量，这多少对网络性能的降低有所贡献，导致网络的拥塞。通过每个概要，将存在关联的一个或多个策略（即应用降低的带宽、流量整形、基于策略的消耗等）。图7.3代表所建议的架构。下面将概述所有架构组件及其功能。

图7.3 用于用户剖析的网络安全架构

该解决方案由两个主要组件组成：分析器（选择采样率和采样模式，并实际上采样流量）和中心式服务器（负责进一步检查被采样的流量，实现预测和客户概要映射）。

7.3.1 分析器

分析器是一种实体，负责流量采样、每条被采样分组的检查，以及将数据分组检查输出插入到数据库中。

这个组件将在第二汇聚路由器（在ISP网络上）中实现。第二汇聚路由器是在数字用户线接入复用器（DSLAM）（见图7.4）之后安装的设备。一般而言，第二汇聚路由器通常称作EDGE路由器，具有大型存储容量和处理能力。分析器由如下实体组成：一个嗅探器、一个采样模块、一个DPI模块和一个数据库，在下面描述。

“嗅探器”负责截获通过路由器的所有流量，并将之传递（复制）到采样模块。“采样模块”负责对由嗅探器交付的流量进行采样，它由三个组件组成：采样模式、采样模式选择和采样率选择。

存在这样一种需要，选择多种采样模式，原因是不存在单一模式，它本身可匹配所有多样化和不可预测的流量变形。采用三种采样模式：基于系统时间的采样、随机采样和分层随机采样，可裁剪最佳采样，这取决于各种条件。

组件“采样模式选择”请求将模式载入直接在上面的组件。将被缴费的模式是使用三个变量计算的：流量总量、路由器计算负载和连接速度。为找出要载入哪种模式，应该使用一种学习算法（如一种神经网络），将依据三个输入变量进行训练。组件“采样率选择”负责找出对流量采样的合适步调，也负责当找到正确的

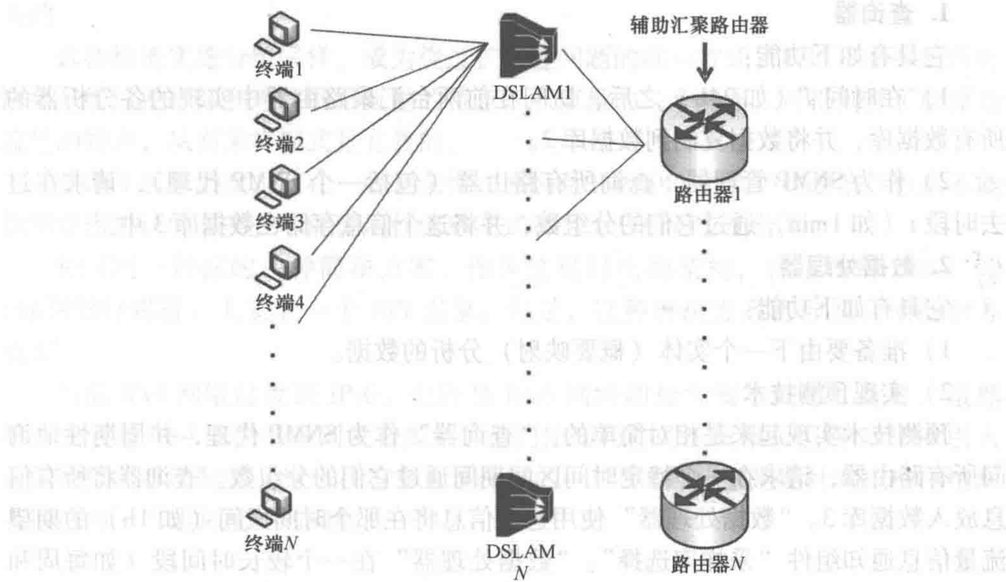


图 7.4　分析器位置

采样率时（使用一项预测技术），对流量实施采样。在采样之后，流量进入下一个模块，即模块 DPI，它负责检查每个被采样的分组并识别它所匹配的技术/协议。DPI 引擎将每条分组检查的结果保存到一个数据库。依据图 7.5 所示的格式，形成该数据库。

	Technology1	Technology2	Technology3	.	.	.	Technology N
Client1	350	357	7	.	.	.	5566
Client2	456	0	456	.	.	.	0
Client3	324545	357	2343	.	.	.	62
Client4	5545	2457	0	.	.	.	27
.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.
Client N	2345667	123246	666	.	.	.	69

图 7.5　数据库 1 格式例

图 7.5 中的数值表示为客户 X 识别出的采用技术 Y［如 BitTorrent、eDonkey、SOPCast、魔兽世界（World of Warcraft）］的分组数。

7.3.2　中心式服务器

一个“中心式服务器”应该负责处理由分析器得到的数据，并实现预测。除了数据库外，中心式服务器将由如下实体组成。

1. 查询器

它具有如下功能：

1）在时间 T（如24h）之后，访问在前两台汇聚路由器中实现的各分析器的所有数据库，并将数据复制到数据库2。

2）作为SNMP管理器，查询所有路由器（包括一个SNMP代理），请求在过去时段 t（如1min）通过它们的分组数，并将这个信息存储在数据库3中。

2. 数据处理器

它具有如下功能：

1）准备要由下一个实体（概要映射）分析的数据。

2）实现预测技术。

预测技术实现起来是相对简单的。"查询器"作为SNMP代理，并周期性地询问所有路由器，请求在一个特定时间区间期间通过它们的分组数。查询器将所有信息放入数据库3。"数据处理器"使用这个信息将在那个时间区间（如1h）的期望流量信息通知组件"采样率选择"。"数据处理器"在一个较长时间段（如每周和每月）结束时重新计算发生在那台路由器中流量总量的统计信息，由此能够以极大的准确度预测在该路由器上的流量总量。

实体"概要映射"是一个神经网络。其输入数据是客户大部分时间使用技术的分组数，或每个客户端技术的总流量百分比，以使神经网络学习要与一个客户相关联的正确的客户概要。

7.4 小结

本章给出有关机制、技术和相关工作的信息，这些涉及下一代IPv6网络的最相关的安全工具。开始时，分析由IPv6网络带来的主要安全问题（相对于IPv4网络）和从IPv4网络过渡到IPv6网络的主要安全弱点。

讨论了下一代IPv6网络将比当前IPv4网络更加开放，这是由于存在大量公开的IPv6地址，几乎可用于我们能够想象的每个联网的设备。这将意味着NAT的终结和新机制的出现，如基于信任的安全，这种机制基于用户的声誉分值和客户剖析。后者可被用来推断临时的非法动作，以及将用户映射到合适的商务缴费模型。这种方法论将支持对大量使用P2P应用的客户进行合适的计费。

因此，本章提出一种可扩展的架构，该架构能够学习每个客户使用网络的使用模式和客户将之用于何种目的（或多种目的）。将客户与一个特定的使用概要相关联。每个用户概要与预定义的策略相关联，这将支持网络的一种自动优化。所给出的架构使用采样方法，我们期望这种方法是有效的，原因是仅在一个24h时段结束时才采样，这是被采样流量评估计算的结果。为了降低这项任务的计算开销，选择采样的相对简单的方式而不是比较复杂的方式，其主要因素之一是与性能问题有

关的。

选择性地实施分组采样，成为绕过扩展性问题的唯一方式，所以不必分析所有网络分组以便感知和理解网络流量的特点。从流量的一个样本，可得到并了解原始流量的特点，从而采样模式是有效的。

在DPI引擎层次，存在一些复杂性，原因是开放DPI也有一些缺陷，包括不能识别加密协议的事实，不能使用任何启发式和行为分析对分组分类。

针对同一目标的一种简单方案，作为这里提出的架构，将是在每个客户端（即调制解调器）上安装一个DPI引擎。但是，这种解决方案将涉及巨大的财务成本。

当前IPv4网络过渡到IPv6，也许是IPv6网络的最大安全问题。此外，虽然IPv6安全与IPv4安全的区别没有多少，它们使用同样的工具集，但确实IPv6引入新的特征，自动配置（这是在本章大量探究的一个专题），这是要针对这些网络深入探讨的一个非常重要的问题。

参考文献

1. L. Ho, C. Macey, and R. Hiller, "A distributed and reliable platform for adaptive anomaly detection in IP networks," *Proceedings of the 10th IFIP/IEEE International Workshop on Distributed Systems: Operations and Management: Active Technologies for Network and Service Management*, 1999.
2. R. Schollmeier, "A definition of peer-to-peer networking for the classification of peer-to peer architectures and applications," *Proceedings of the First International Conference on Peer-to-Peer Computing*, IEEE, 2002.
3. D. Awduche, "MPLS and traffic engineering in IP networks," *IEEE Communications Magazine*, 1999.
4. R. Jain, "Congestion control in computer networks: issues and trends," *IEEE Network Magazine*, 1990.
5. V. Jacobson and M. Karels, "Congestion avoidance and control," ACM SIGCOMM, Symposium on Communication Architectures and Protocols, 1988.
6. N. Collignon, "IPv6 network security threats," In a technical report by Herve Schauer Consultants, 2006.
7. S. Blake, D. Black, M. Carlson, E. Davies, Z. Wang, and W. Weiss, "An architecture for differentiated services," IETF RFC 2475. 1998.
8. J. Steffen, "Sistemas de Detecção de Intrusão," Monografia (Bacharelato em Ciência da Computação), Instituto de Ciências Exactas e Tecnológicas, 2003.

9. N. Desai, "Intrusion prevention systems: the next step in the evolution of IDS," *Security Focus*, 2003. Disponível em: http://www.securityfocus.com/print/infocus/1670 Acedido em: Abril de 2010.
10. M. Rash and A. Orebaugh, "Intrusion prevention and active response: deploying network and host IPS," 2005.
11. L. Carvalho, "Segurança de Redes," 2006.
12. S. Northcutt, L. Zeltser, S. Winters, K. Fredrick, and R. Ritchey, *Inside Network Perimeter Security: The Definitive Guide to Firewalls, VPNs, Routers, and Intrusion Detection Systems*, New Riders, 2003.
13. A. Marcelo and M. Pitanga, "Honeypots: a Arte de Iludir hackers," 2003.
14. M. Reis, "Forense computacional e sua aplicação em segurança imunológica," MsC thesis, Instituto de Computação, Universidade Estadual de Campinas, 2003.
15. J. Beale, J. C. Foster, J. Posluns, R. Russell, and B. Caswell, "Snort 2.0 intrusion detection," *Syngress*, 2003.
16. K. Mccloghrie, S. Robert, J. Walrand, and A. Bierman, "Sampling packets for network monitoring" United States Patent 6920112, 2005.
17. D. Harrington, R. Presuhn, and B. Wijnen, "An architecture for describing simple network management protocol (SNMP) management frameworks," IETF RFC 3411, 2002.
18. R. Presuhn, J. Case, K, McCloghrie, M. Rose, and S. Waldbusser, "Version 2 of the protocol operations for the simple network management protocol (SNMP)," IETF RFC 3416, 2002.
19. A. Leinwand and K. Conroy, *Network Management—A Practical Perspective*, Addison-Wesley, 1996.
20. W. Stallings, *SNMP, SNMPv2 and RMON—Practical Network Management*, Addison-Wesley, 1996.
21. "Cisco Systems Inc. NetFlow services and applications," White paper, 2007.
22. P. Phaal, S. Panchen, and N. McKee, "InMon Corporation's flow: a method for monitoring traffic in switched and routed networks," IETF RFC 3176, 2001.
23. M. Fullmer and S. Romig, "The OSU flow-tools package and CISCO netflow logs," *Proceedings of the Fourteenth Systems Administration Conference (LISA-00)*, 2000.
24. P. Phaal and S. Panchen, "Packet sampling basics," 2007.
25. B. Choi, J. Park, and Z. Zhang, "Adaptive packet sampling for accurate and scalable flow measurement," *Proceedings of the IEEE Globecom*, 2004.
26. "Technical white paper for NetStream," Report available on May 2010 at www.huawei.com/products/datacomm/pdf/view.do?f=65.
27. N. Duffield, C. Lund, and M. Thorup, "Estimating flow distributions from sampled flow statistics," *Proceedings of the ACM SIGCOMM 2003*,

2003.

28. C. Estan and G. Varghese. "New directions in traffic measurement and accounting," *Proceedings of the ACM SIGCOMM 2002*, 2002.
29. N. Hohn and D. Veitch, "Inverting sampled traffic," ACM Internet Measurement Conference—IMC'03, 2003
30. G. Silvestre, C. Kamienski, S. Fernandes, and D. Sadok, "Análise Quantitativa e Qualitativa de Tráfego P2P baseada na Carga Útil dos Pacotes," 2004.
31. N. Duffield, "Sampling for passive Internet measurement: a review," *Statistical Science*, 2004.
32. S. Fernandes, T. Correia, C. Kamienski, D. Sadok, and A. Karmouch, "Estimating properties of flow statistics using bootstrap," IEEE MASCOTS 2004, 2004.
33. N. Duffield, C. Lund, and M. Thorup, "Charging from sampled network usage," *Proceedings of the ACM SIGCOMM Internet Measurement Workshop*, 2001.
34. C. Kamienski, T. Souza, S. Fernandes, G. Silvestre, and D. Sadok, "Caracterizando Propriedades Essenciais do Tráfego de Redes através de Técnicas de Amostragem Estratificada," Maio 2005.
35. T. Zseby, M. Molina, F. Raspall, N. Duffield, and S. Niccolini, "Sampling and filtering techniques for IP packet selection," IETF RFC 5475, 2009.
36. N. Duffield, D. Chiou, B. Claise, A. Greenberg, M. Grossglauser, and J. Rexford, "A framework for packet selection and reporting," IETF RFC 5474, 2009.
37. S. Fernandes, C. Kamienski, D. Mariz, and D. Sadok, "Avaliação de Técnicas de Agrupamento na Amostragem de Tráfego na Internet," 24th Brazilian Symposium on Computer Networks (SBRC 2006), 2006.
38. S. Fernandes, C. Kamienski, D. Mariz, and D. Sadok, and J. Kelner, "A stratified traffic sampling methodology for seeing the big picture," *International Journal of Computer and Telecommunications Networking*. 52(14), Elsevier North-Holland, New York, NY, 2008.
39. R. Jurga and M. Hulbój, "Packet sampling for network monitoring," Technical report, 2007.
40. T. Zseby, "Stratification strategies for sampling-based non-intrusive measurements of one-way delay," *Passive and Active Measurement Workshop Proceedings*, 2003.
41. V. Paxson, G. Almes, J. Mahdavi, and M. Mathis, "Framework for IP performance metrics," IETF RFC 2330, 1998.
42. S. Chaudhuri, R. Motwani, and V. Narasayya. "On random sampling over joins," *SIGMOD 99: Proceedings of the 1999 ACM SIGMOD International Conference on Management of Data*, 1999.

43. N. Duffield, C. Lund, and M. Thorup, "Learn more, sample less: control of volume and variance in network measurement," *IEEE Transactions in Information Theory*, 2005.

44. G. Manku and R. Motwani. "Approximate frequency counts over data streams," *Proceedings of the 28th International Conference on Very Large Databases*, 2002.

45. H. Chen, F. You, X. Zhou, and C. Wang, "The study of DPI identification technology based on sampling," *Information Engineering and Computer Science*, 2009.

第8章 物 联 网

Syam Madanapalli

下一波通信发生在移动互联网上物体（电子交换机、灯泡、门锁、家庭仪器、工业机械和其他物体）之间，用于节省能量和资源，并用于远程监测、控制和管理，同时使人们的生活更美好。要为互联的这种物体的数量将是数十亿到数兆亿的，这使现存的因特网显得微不足道。本章介绍物联网及其网络架构、协议栈和应用，也描述在实现物联网中需要因特网协议版本6（IPv6）。

8.1 物联网：新型因特网

8.1.1 引言

因特网是人类曾创造的最成功的、新颖的和大规模的网络。在20世纪90年代，人们开始使用因特网共享信息和知识，其中使用到万维网，这一般被看作第一代因特网。正在进行的第二代因特网是社交网络的平台。人们期望，下一次革命，第三代因特网，将是地球上每个可能物体的互联，这将产生一个新的因特网，称作物联网。

物联网是一个自组织和自愈物体组成的网络，以因特网作为机器间、机器-人之间以及人-机器之间交换信息的主要通信媒介。

物联网应用包括各种远程监测和控制应用，包括但不限于，连接的家庭（智能家庭）、智能电网、工业自动化、舰队管理、资产跟踪、农业应用、航空和网络战应用。

物联网展现四个趋势，这带来挑战，但要得到巨大优势则要面对挑战。这些有：

1）规模：连接到因特网的节点（设备/物体/传感器）数正在增长并将从数十亿增长到数兆亿。

2）异构性：节点的类型、连接的类型以及各种信息和应用，在数量方面都在增长，并就互操作能力方面施加极大的挑战。

3）水平化（Horizontalization）：物联网中的节点可参与到多项应用，并不绑定到一项特定服务。这使物联网成为一个平台，并为开发各种应用提供一个巨大的机会。

4）移动性：物体越来越多地采用无线连接。而且这些节点中的一些节点可被

附接到由移动实体承载的物体。

IPv6[3-7]，因特网的下一代协议和因特网协议版本4（IPv4）[8]的后继者，可处理这些趋势，方法是为各种类型的数兆设备和物体提供到移动因特网的统一连接能力。在实现成本有效解决方案中有所帮助的其他技术有短距离、低速率的无线技术［如美国电气电子工程师学会（IEEE）802.15.4[1]］，射频识别（RFID），传感器，移动设备（移动电话便签本等）和实时万维网。

8.1.2 社会影响

人们期望，物联网应用的范围将带来如今社会中生活模式的变化，方法是提高人们生活的质量。例如，物联网可被用于如下方面：

1）在家庭或医院中使用无线传感器，实施患者和老年人生命特征的监测，提供改进的监测准确度，同时对患者也是比较方便的。

2）降低森林被砍伐程度，方法是为树木装备传感器，这可向地方权威机构提供实时信息。

3）连接的车辆，将有助于降低交通堵塞并改进它们的再循环能力，由此将降低它们的碳排放。

人们期望物联网会放大大型联网通信对社会的深远影响，逐步地导致真正的生活方式变化。物联网也可改进公民的生活质量，为工人交付新的和更好的工作，带来商务机遇和工业方面的增长。

8.2 物联网的特点

物联网典型地由因特网上互联的低功率无线个域网（LoWPAN）组成。但是，可能存在许多其他物理媒介，如以太网、电力线通信（PLC）、IEEE 802.11、蜂窝服务［第二代（2G）/第三代（3G）/长期演进（LTE）］和全球微波接入互操作性（WiMAX），这些可被机器用来进行通信。本章将焦点放在无线技术（特别是802.15.4）作为节点之间信息传递的媒介。一个LoWPAN典型地由高度受约束节点［受限的中央处理单元（CPU）、内存、电池］以低速率、低功耗和丢失性无线电链路（典型的是IEEE 802.15.4）互联组成。本节探索LoWPAN节点的典型特点和LoWPAN的各种考虑。

8.2.1 典型的LoWPAN节点的特点

1. 受限的处理能力

典型的LoWPAN节点有8/16比特微型控制器，CPU速度在10MHz左右。要求更大处理能力的一些节点也许有32比特的内核（典型的ARM7），CPU速度在数十兆赫量级。

2. 小型内存容量

LoWPAN 设备的常见随机访问内存（RAM）尺寸为数千字节，通常为 8KB。但是，存在各种 RAM 尺寸，从 1KB 到 256KB。

3. 小型面积

典型的只读内存（ROM）尺寸从 48KB 到 128KB，可容纳非常少量的代码。

4. 低功率

LoWPAN 中的各节点正常情况下是由电池驱动的。其无线电经常有大约 10 ~ 30mA 的电流输出，这取决于所用的传输功率水平。为达到高达 30m 的常见室内距离和高达 100m 的室外距离，所用传输功率被设置在 0 ~ 3dBm。当切换到睡眠模式时，CPU 功率消耗经常降低 1000 倍。在 LoWPAN 中典型占空比小于 0.1%，而且对于一些应用，电池会被焊接到节点上，并持续到设备的整个寿命期间。

5. 短距离

由 IEEE 802.15.4 定义的个人操作空间（POS）意味着 10m 的范围。对于真正的实现，LoWPAN 无线电的范围典型地是以十几米来度量的，并在视距通信中可达 100m 以上。

6. 低比特率

IEEE 802.15.4 标准定义了 20kbit/s、40kbit/s、100kbit/s 和 250kbit/s（250kbit/s 最普遍用在当前部署中）的空中数据率。但是，所要求的实际数据率要低得多，在大部分时间段，设备典型地处在睡眠模式。

8.2.2 LoWPAN

一般情况下，一个 LoWPAN 由 LoWPAN 主机和 LoWPAN 路由器（或 LoWPAN 网状节点）组成，所有这些都被称作 LoWPAN 节点。LoWPAN 主机可以是信息源或信息目的地，且 LoWPAN 网状节点/路由器是典型的 LoWPAN 主机，它在源-目的对之间转发数据。LoWPAN 路由器和 LoWPAN 网状节点之间的区别是它们所工作的层。LoWPAN 路由器实施 IP 路由，LoWPAN 网状节点工作在链路层上面，并使用链路层地址进行转发和多芯片功能。

一个典型的 LoWPAN 拓扑如图 8.1 所示。到 LoWPAN 之外通信节点的通信，对于方便的数据收集和远程控制目的，正变得日渐重要。网关或边缘路由器被用来将一个 LoWPAN 互联到其他网络，或通过连接多个 LoWPAN，形成一个扩展的 LoWPAN。

典型的 LoWPAN 考虑以下内容。

1. 部署

LoWPAN 节点可随机地散布，或它们可在一个 LoWPAN 中以一种有组织的方式加以部署。LoWPAN 节点可以增量方式加以部署。一些节点可能需要无缝地去除或替换。

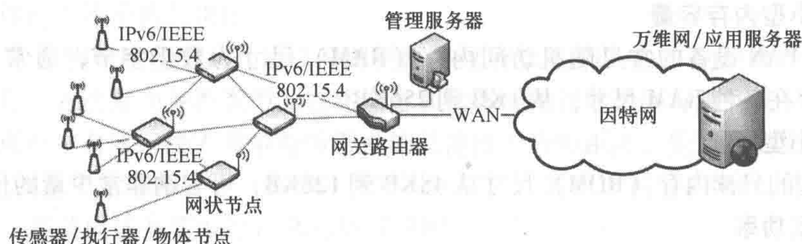


图 8.1 连接到因特网的典型 LoWPAN

一个 LoWPAN 中的各节点可以工作在星形或网状拓扑，这取决于应用需求。

2. 网络尺寸

网络尺寸典型地是以 LoWPAN 节点数度量的，要求有这些节点才能覆盖想要的应用。在一个 LoWPAN 中涉及的节点数可以是少量的（10 个节点）、中等（几百个）或大量的（超过 1000 个）。

3. 电源

LoWPAN 节点可从电池供电或主供电汲取能量。一个典型的 LoWPAN 由混合节点组成，一些是电池供电的，其他节点是主供电的。电源可以是从太阳电池、振动能或其他能量源得到的。

4. 连接能力

一个 LoWPAN 中的无线链路是低比特率的和丢失性的，这是由于外部因素（如极端环境、移动性）或编程的占空周期（如睡眠模式）。所以，网络连接能力可来自间歇性的（正常的断开）到不定时发生的（几乎总是断开的网络）。

5. 多跳通信

单跳对简单的星形拓扑是足够的，但对于更复杂的拓扑（如网状或树）则要求多跳通信方案。

6. 流量模式

几种流量模式可用于一个 LoWPAN 中，这取决于应用需要，包括点到多点、多点到点和点到点模式。

7. 占空周期

电池供电的 LoWPAN 节点要求较小的占空周期，以便降低功率消耗。这些节点将花费其大部分工作时间在睡眠模式。但是，每个设备周期性地侦听射频信道，以便确定一条消息是否在发送中。网络设计人员应该确定电池消耗和消息延迟之间的平衡。

8. 安全性

LoWPAN 可携带敏感信息，并要求高级安全性支持，其中信息的可用性、完整性和机密性是至关重要的。在智能电网、患者的健康监测和其他任务关键性应用的情形中，需要这种高级的安全性。

9. 移动性

移动性是固有地存在于 LoWPAN 的无线特点。一个 LoWPAN 中的各节点可到处移动或被移动。移动性可以是一项诱发因素（如一辆机动车中的传感器），所以是不可预测的或一个受控的特点（例如，在一个供应链中提前计划的运动）。

10. 服务质量

对于任务关键性的应用，以实时地满足合适服务质量（QoS）进行信息传播是一项重要特征。QoS 是 LoWPAN 实现的一个挑战，实现是由资源受约束的节点组成的。

8.3 实现物联网的标准

物联网的应用将有各种细分市场的广度和深度，所以将有数千种不同类型的设备：

1）不兼容的硬件配置。

2）不兼容的 CPU、架构和内存面积。

3）不同类型的操作系统，专用的和开源的（和没有操作系统的一些设备）。

4）不同类型的竞争性的连接标准（以太网、ZigBee㊀、HomePlug、IEEE 802.15.4、Wi-Fi、PLC 等）。

依据 Harbor 研究公司的泛在因特网/M2M 预测报告（2009）[16]，智能设备发货的数量将从 2008 年的 7300 万件增长到 2013 年的 43000 万件。存在大量不兼容的设备，这种多样性将随着新的芯片/操作系统/连接性进入市场而进一步增加。这种多样性需要针对物体间的透明通信进行桥接处理。传输控制协议/因特网协议（TCP/IP）协议族证明了在任何种类媒介（以太网、蜂窝、光传输等）之上传递任何信息（数据、话音、视频、实时信息）的这样一种能力（见图 8.2）。作为 IPv4 的后继者，IPv6 具有巨大的寻址能力，可无缝地连接具有任何种类多样性的数兆亿设备。

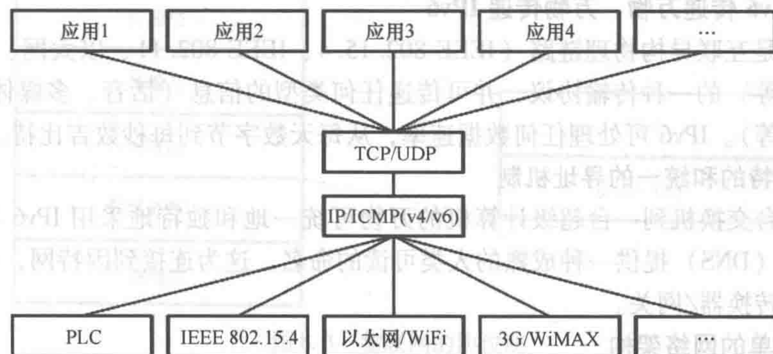


图 8.2 IP 传递万物，万物传递 IP

㊀ ZigBee 是低成本、低功率、无线网状联网的另一种标准。

但是，IPv6具有一个较大的首部（40字节），最小分组尺寸1280字节可高效地在IEEE 802.15.4之上传输，后者的净荷是127字节。LoWPAN传递IPv6（6LoWPAN）[13,14]，是由因特网工程任务组（IETF）（为因特网开发标准）制定的一个开放标准，是在IEEE 802.15.4之上传递IPv6，在IPv6网络层和数据链路层之间引入一个适配层。6LoWPAN使IPv6在IEEE 802.15.4之上高效传输，方法是提供如下关键功能，这将在8.7节比较详细地加以讨论。

1）首部压缩：通过假定在一个LoWPAN内使用常用值，压缩IPv6首部字段。当适配层可从802.15.4帧携带的链路层信息中推断出首部字段或基于共享语境的简单假定时，从一个分组中可删去这些首部字段。

2）分片：IPv6分组被分片成多个链路层帧，以便处理1280字节的IPv6最小最大传输单元（MTU）需求。

3）网状路由：为支持IPv6数据报的层2转发，适配层可为一个IP跳的端点携带链路层地址。另外，IP栈可通过层3转发完成个域网内（PAN内）路由，其中每个802.15.4无线电跳是一个IP跳。前者被称作网下（mesh-under）路由，后者称作网上（mesh-over）路由。

下面介绍IPv6的角色及因特网的角色。

IPv6及因特网用作物体通信的媒介具有许多优势，也许是构建一个可扩展物联网的唯一解决方案。

1. 开放标准

全IP网络基于真正的开放标准，设施和网络建造者可在世界间混合和匹配来自多个厂商的设备。开放标准也带来低拥有成本的优势。物联网正在为一个可预见的未来而不确定地演进着，基于IPv6及因特网的通信基础设施有助于快速应用创新和新功能的交付，这在今天是不可能的甚至不可想象的。

2. IPv6传递万物，万物传递IPv6

IPv6是互联异构物理链路（IEEE 802.15.4、IEEE 802.11、以太网、WiMAX、蜂窝网络等）的一种传输协议，并可传递任何类型的信息（话音、多媒体、数据、实时信息等）。IPv6可处理任何数据速率，从每天数字节到每秒数吉比特。

3. 独特的和统一的寻址机制

从一台交换机到一台超级计算机的万物可统一地和独特地采用IPv6寻址，而域名服务（DNS）提供一种成熟的人类可读的命名。这为连接到因特网，去除了地址和协议转换器/网关。

4. 简单的网络架构

每台设备都使用IP（没有协议转换器、没有地址转换器、没有信息转换器）产生一个简单的全IP网络，这是容易运作和维护的，并具有低的拥有成本。

5. 无缝web服务

因特网上最成功的应用是万维网，因特网架构提供了web服务的无缝集成。适

配 IPv6 的做法，将以如今丰富的资源、工具、技术和模型支持新颖应用的开发。同样，应用可独立于传输网络进行开发，原因是应用没有绑定到设备，并可从开放的市场购买到。

6. 端到端安全

IETF 为与 IP 一起使用开放了大量安全协议，它们在多个异构的、互联的管理域间提供端到端安全。在没有重新发明车轮的情况下，这些协议和知识可被重用于构建物联网。

7. 现有的资源和知识

因特网已经构建了 20 年以上（商业化和私人运行因特网服务引入开始于 20 世纪 80 年代），存在巨量工具和技术（被开发和成功地部署）；为构建物联网，而重用它们，得到一个鲁棒的、容易运作和维护的网络，不要求任何其他经过培训的人员。

上述优势使 IPv6 成为构建物联网的明确选择。

8.4　用于物联网的协议层

像任何其他连接的设备一样，物联网由典型的通信协议层组成，如图 8.3 所示。LoWPAN 的网络层由 IPv6 和称作 6LoWPAN 的一个子层组成，后者将 IPv6 传输适配到 IEEE 802.15.4，它提供物理（PHY）和数据链路层的功能。典型情况下，用户数据报（UDP）被用作传输层。应用层功能由各种标准开发组织定义，著名的有 ZigBee、W3C 和 IETF。为物联网使用 IP 的整个目标是无缝地与因特网上的其他设备通信。因特网上的绝大多数设备使用可表示状态转换（REST）架构通信，所以物联网也将在近期基于 RESTful 架构上。

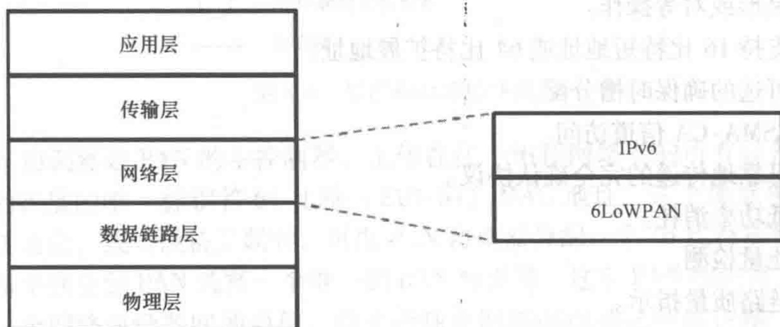


图 8.3　物联网的协议层

8.5　用于物联网的 IEEE 802.15.4——PHY 和 MAC

IEEE 802.15.4 是 LoWPAN 实现的主导无线电标准，它规范 PHY 和媒介访问

控制（MAC）子层。IEEE 802.15.4规范低功率无线电（典型地由电池供电），它适合于短距离通信，典型地从10m到100m。IEEE 802.15.4 MAC基于带有冲突避免的载波侦听多路访问（CSMA-CA）技术，这可在IEEE 802.11标准中找到。MAC支持星形和对等网络拓扑。IEEE 802.15.4设备竞争媒介访问，但是，IEEE 802.15.4也使用超帧结构为时槽访问时间关键数据规范一个可选机制。

IEEE 802.15.4的主导频带有868MHz、915MHz和2.45GHz。

8.5.1　868/915MHz频带

这个频带支持三种不同的PHY实现：

1）直接序列扩频（DSSS）PHY，采用二相移相键控（BPSK）调制。

2）DSSS PHY采用偏移四相移相键控（O-QPSK）调制。

3）平行序列扩频（PSSS）PHY，采用BPSK和移幅键控（ASK）调制。

它支持的数据速率有20kbit/s、40kbit/s，可选地有100kbit/s和250kbit/s。

在915MHz频带支持30个信道，在868MHz频带支持3个信道。

8.5.2　2.45GHz ISM频带

PHY基于DSSS，采用O-QPSK调制。它支持250kbit/s的数据速率。它支持16个信道。

IEEE 802.15.4的设计，是为支持容易安装、可靠的数据传递、短距离操作、极低成本和合理的电池寿命，同时维护一个简单的和灵活的协议。IEEE 802.15.4的一些特点如下：

1）空中数据速率有250kbit/s、100kbit/s、40kbit/s和20kbit/s。

2）星形或对等操作。

3）支持16比特短地址或64比特扩展地址。

4）可选的确保时槽分配。

5）CSMA-CA信道访问。

6）可靠地传递的完全确认协议。

7）低功率消耗。

8）能量检测。

9）链路质量指示。

IEEE 802.15.4规范两种不同的设备类型：完全功能的设备（FFD）和精简功能的设备（RFD）。

FFD特点如下：

1）可工作在三种模式，用作一个PAN协调器、一个协调器或一台设备。

2）与任何其他设备通信。

3）实现完全的协议集。

4）是为网状路由应用设计的。

RFD特点如下：

1）受限于星形拓扑或一个对等网络中的端设备。

2）不能成为一个PAN协调器。

3）是一个精简的协议集，带有最小面积和低成本的简单实现。

4）是为极简单的应用设计的，如一台轻量交换机；它们不需要发送大量数据，并具有非常低的占空周期（<0.1%）。

8.5.3 网络拓扑

在相同物理信道上通信的一个个人工作空间内的两台或多台IEEE 802.15.4，构成一个LoWPAN，这个LoWPAN必须至少包括一个FFD，作为中央受控的，称作PAN协调器。取决于应用需求，一个IEEE 802.15.4网络可工作在两种拓扑中，即星形拓扑或对等拓扑。这两种拓扑如图8.4所示。

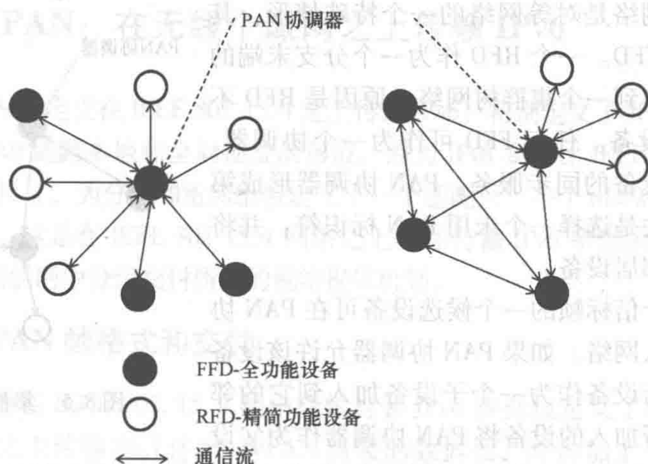


图8.4 星形和对等拓扑图示

PAN协调器是PAN的主控制器。工作在任一拓扑网络上的所有设备将具有一个独特的扩展的唯一标识符64比特（EUI-64）MAC地址。这个地址可用于PAN内的直接通信，或当设备关联时，可由PAN协调器分配一个16比特短地址并加以使用。每个独立的PAN选择一个唯一的PAN标识符。这个PAN标识符支持使用短地址的一个网络内设备间的通信，并支持独立网络间设备之间的传输。除了管理PAN外，PAN协调器也有一个特定的应用，它经常是采用主供电方式的。

1. 星形网络拓扑

在星形拓扑中，通信是在设备和PAN协调器之间建立的。一个星形网络的基本结构如图8.4所示。在一个FFD上电之后，它可建立自己的网络，并成为PAN协调器。所有星形网络独立于当前工作的所有其他星形网络而工作。通过选择当前

不受影响的无线电球形空间内任何其他网络所用的一个 PAN 标识符，做到这一点。一旦选中 PAN 标识符，PAN 协调器支持其他设备（潜在地有 FFD 和 RFD）加入它的网络。星形拓扑的典型应用有家庭自动化、个人计算机（PC）外设、玩具和游戏以及个人保健。

2. 对等网络拓扑

对等拓扑也有一个 PAN 协调器。但是，它不同于星形拓扑的是，任意设备可与任何其他设备通信，条件是只要它们在相互的通信范围内。对等拓扑支持实现更复杂的网络形成，如网状联网拓扑。诸如工业控制和监测、无线传感器网络、资产和库存跟踪、智能农业以及安全等应用，将受益于这样一种网络拓扑。一个对等网络可以是独特的、自组织的和自愈的。它也支持多跳将消息从任意设备路由到网络上的任何其他设备，其中使用网状路由技术，这典型地在高层协议的帮助下做到这一点。

对等通信拓扑的一个例子是集群树，如图 8.5 所示。集群树网络是对等网络的一个特殊情形，其中多数设备是 FFD。一个 RFD 作为一个分支末端的一个叶设备连接到一个集群树网络，原因是 RFD 不支持关联其他设备。任何 FFD 可作为一个协调器，并提供到其他设备的同步服务。PAN 协调器形成第一个集群，方法是选择一个未用 PAN 标识符，并将信标帧广播到邻居设备。

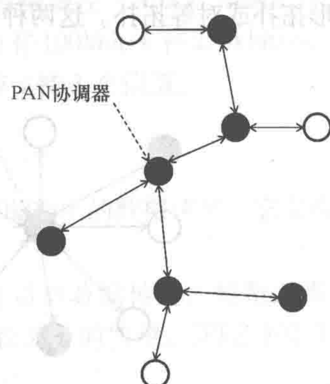


图 8.5 集群树拓扑

接收到一个信标帧的一个候选设备可在 PAN 协调器处请求加入网络。如果 PAN 协调器允许该设备加入，则它将新设备作为一个子设备加入到它的邻居列表。之后新加入的设备将 PAN 协调器作为父设备加入到它的邻居列表，并开始传输周期性的信标，之后其他后续设备可在那台设备处加入网络。

8.6 IPv6

将在 IEEE 802.15 MAC 数据帧上携带 IPv6 分组，典型情况下，建议 IPv6 分组在请求确认的帧中加以携带。一个 IEEE 802.15.4 PAN 被处理为 IPv6 操作的单一 IPv6 链路。

在 IEEE 802.15.4 上传输 IPv6 分组的 MTU 尺寸是 1280 字节，这是最小 IPv6 MTU。但是，一个完整的 IPv6 分组不能装在一个 IEEE 802.15.4 帧中。IEEE 802.15.4 协议数据单元有不同尺寸，取决于存在多少开销。

最大物理层分组尺寸是 127 字节。

IEEE 802.15.4 中的最大帧是 25 字节。

最大链路层安全（AES-CCM-128 情形）开销是 21 字节。

可用于 IPv6 分组的净荷尺寸是 81（127－25－21）字节。

考虑一个 40 字节 IPv6 首部和一个 8 字节 UDP 首部，可用于应用的实际净荷尺寸是 33 字节。

上述考虑使 IPv6 在其原始形式中运行在 IEEE 802.15.4 之上是非常低效的，并导致如下观察结果：

1）必须提供适配层，满足 IPv6 对一个最小 MTU 的需求。但是，人们期望 IEEE 802.15.4 的多数应用将不适用这种大型的分组，且小型应用净荷与合适的首部压缩一起使用，将产生适合在单个 IEEE 802.15.4 帧内的分组。

2）即使上述空间计算给出最坏场景，但它确实指出这样的事实，即对于 IPv6 在 IEEE 802.15.4 之上传输，首部压缩是紧迫要求，是不可避免的。

8.7 6LoWPAN：在无线个域网之上传输 IPv6

一个 6LoWPAN 定义在 IEEE 802.15.4 之上传输 IPv6，特别定义了 IPv6 分组传输的帧格式，以及 IPv6 链路本地和全局地址的形成。因为 IPv6 要求比 IEEE 802.15.4 帧尺寸大得多的分组尺寸，为分片和重新组装定义了一个适配层。一个 6LoWPAN 也为首部压缩定义了机制，这是在 IEEE 802.15.4 网络之上实际传输 IPv6 所需的，也定义了在 IEEE 802.15.4 网状网中分组交付所需的就绪提供机制。

8.7.1 LoWPAN 帧格式和交付

6LoWPAN 为在 IEEE 802.15.4 帧中高效封装 IPv6 数据报定义了各种首部。在 IEEE 802.15.4 之上传输的所有 6LoWPAN 封装的数据报，外部加上一个封装首部栈。首部栈每个首部包含一个首部类型，后跟零个或多个首部字段。6LoWPAN 首部的类型如图 8.6 所示。

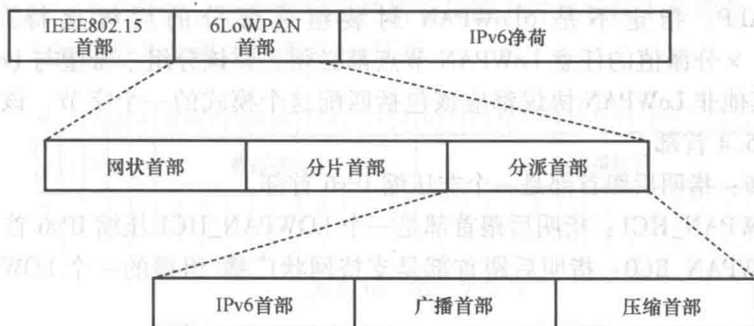


图 8.6 6LoWPAN 首部

不是所有首部在所有时间都存在。如果存在，它们必须按如下顺序出现：

1）网状寻址首部。

2）广播分派首部。

3）分片首部。

4）压缩（HC1）首部。

所有协议数据报（IPv6，压缩 IPv6 首部等）必须在前面有有效 6LoWPAN 封装首部之一。

1. 6LoWPAN 分派首部

分派首部的格式如图 8.7 所示。

分派类型 (2比特)	分派 (6比特)	分派特定的首部

图 8.7　6LoWPAN 分派首部

分派首部的字段定义如下：

1）分派类型：0 比特被定义为第一比特，1 比特被定义为第二比特。

2）分派：是一个 6 比特选择器，识别直接后跟分派首部的首部类型。

3）分派特定的首部：是由分派首部确定的一个首部。

图 8.8 给出了为一个 6LoWPAN 定义的分派首部列表。

分派首部 (前8比特)	描述
00 xxxxxx	NALP：不是一个LoWPAN帧
01 000001	IPv6：非压缩IPv6地址
01 000010	LOWPAN_HC1：LOWPAN_HC1压缩的IPv6
01 010000	LOWPAN_BC0：LOWPAN_BC0广播
01 111111	ESC:后跟其他分派字节
所有其他值	保留的：为未来用途保留

图 8.8　分派值比特模式

1）NALP：指定不是 6LoWPAN 封装组成部分的后续比特，遇到 00 ××××××分派值的任意 LoWPAN 节点都必须丢弃该分组。希望与 LoWPAN 节点共存的其他非 LoWPAN 协议都应该包括匹配这个模式的一个字节，该字节直接后跟 802.15.4 首部。

2）IPv6：指明后跟首部是一个非压缩 IPv6 首部。

3）LOWPAN_HC1：指明后跟首部是一个 LOWPAN_HC1 压缩 IPv6 首部。

4）LOWPAN_BC0：指明后跟首部是支持网状广播/组播的一个 LOWPAN_BC0 首部。

5）ESC：指明后跟首部是分派值的单个 8 比特字段。它支持大于 127 的分派值。

2. 网状寻址类型和首部

网状类型由1和0定义为前两个比特。网状类型首部如图8.9所示。



图8.9 网状寻址类型和首部

字段定义如下：

1）V：一个1比特字段。如果源发［或"恰好是第一个"（very first）］地址是一个IEEE EUI-64地址，则为0；如果它是一个短16比特地址，则为1。

2）F：一个1比特字段。如果最终目的地址是一个IEEE EUI-64地址，则为0；如果它是一个短16比特地址，则为1。

3）剩下的跳数：一个4比特字段。在将这条分组往下一跳发送之前，每个转发节点将这个值减1。如果剩下的跳数减少到0，则分组不再进一步转发。值0×F被保留，并指明直接后跟的一个8比特剩余的深度跳字段，支持一个源节点指明大于14跳的一个跳限制。

4）源发者地址：源发者的链路层地址。

5）最终目的地址：最终目的地的链路层地址。

注意，V和F比特支持16比特和64比特MAC地址构成的一个混合地址。这是有用的，至少支持网状层"广播"，原因是802.15.4广播地址被定义为16比特短地址。

3. 分片首部

如果整个净荷（IPv6数据报）可封装在单个802.15.4帧内，则它不分片，且LoWPAN封装不应包含一个分片首部。如果数据报不能封装在单个IEEE 802.15.4帧内，则它必须被分解为多个链路分片。分片偏移是以8字节的整数倍表示的，所以除了最后一个分片外，一个数据报的所有链路分片在长度上必须是8字节的整数倍。第一个链路分片的格式如图8.10所示。



图8.10 第一个分片

第二个和后续的链路分片（直到最后一个分片并包括该分片）必须包含一个

分片首部，它符合如图 8.11 所示的格式。

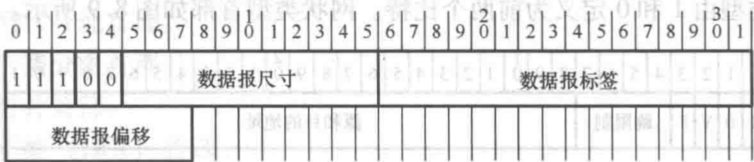


图 8.11 第二个和后续分片

分片首部字段定义如下：

1）数据报尺寸：这个 11 比特字段对链路层（6LoWPAN）分片之前（但在 IP 层分片之前，如果有的话）的整个 IP 分组尺寸进行编码。对于一个 IPv6 分组的所有链路层分片，数据报尺寸的值必须是一样的。

2）数据报标签：这个字段是 16 比特长，且初始值是没有确定的，一台设备可检取一个随机初始值。对于一个净荷（如 IPv6）数据报的所有链路分片，数据报标签的值必须是相同的。对于后续的被分片数据报，发送方必须增加数据报标签的值。数据报标签被增加的值必须从 65535 环回到 0。

3）数据报偏移：这个字段仅存在于第二个和后续链路分片中，必须指明偏移，是以 8 字节的增量表示的，指从净荷数据报的开始部分的分片偏移。这个字段有 8 比特长。

链路分片的接收方将使用如下信息，重构 IPv6 数据报：

1）发送方的 802.15.4 源地址（或在一个网状配置中源发者的地址）。

2）目的地的 802.15.4 地址（或在一个网状配置中的最终目的地地址）。

3）数据报尺寸。

4）数据报标签，识别属于一个给定数据报的所有链路分片。

在接收到一个链路分片时，接收方开始构造原始未分片分组，其尺寸是数据报尺寸。它使用数据报偏移字段确定在原始未分片分组内各分片的位置。当一个节点第一次接收到带有一个给定数据报标签的一个分片时，它启动一个重组定时器。当这个定时器超时时，如果整个分组没有被重新组装，则丢弃已存在的分片。重新组装超时被设置为 60s 的最大值（这也是 IPv6 重新组装规程[3]中的超时）。

8.7.2 一个 6LoWPAN 的邻居发现

IPv6 邻居发现（ND）[6]是 IPv6 主机发现在线（on-link）路由器、前缀和其他配置信息的一个信令协议。它也为地址解析和重复地址检测（DAD）指定机制。

因为 LoWPAN 无线电链路是丢失性的，展示出间歇性连接，所以 6LoWPAN 的 IPv6 ND 必须就如下方面做出优化：

1）通过避免使用组播洪泛并降低使用链路范围组播消息，以此最小化信令。

2）优化主机及其缺省路由器之间的接口。

3）支持正处于睡眠的主机。

4）最小化节点的复杂性。

5）依据首部压缩技术所需，将语境信息传播到各主机。

一个6LoWPAN的ND优化适用于网状网之下（mesh-under）和路由之上（route-over）配置。在网状网下的配置中，仅存在6LoWPAN边界路由器和主机；在网状网下拓扑中，不存在6LoWPAN路由器。

优化的最重要部分是演进的主机到路由器交互通信，这支持睡眠中的节点，并避免使用组播ND消息，例外情况是一台主机找到缺省路由器的一个初始集合并重新做出这种决定的情形，此时那些路由器集已经不可达。

对IPv6邻居发现的扩展（RFC 4861）

6LoWPAN ND指定对IPv6 ND的如下优化和扩展[6]。

1）路由器通告信息的主机初始刷新。这去除了从主机到主机的周期性的或非请求路由器通告的需求。

2）如果使用基于EUI-64的IPv6地址，则不需要DAD。

3）如果使用DHCPv6[17]来指派地址，则DAD是可选的。

4）一种使用主机和路由器之间新的地址注册选项的新地址注册机制。这消除了路由器使用组播邻居请求来发现主机并支持睡眠中主机的需要。这也支持相同的IPv6地址前缀用在一个路由之上6LoWPAN间。它为DAD提供主机到路由器接口。

5）一个新的可选路由器通告选项，用于6LoWPAN首部压缩所用的语境。

6）一个新的可选机制，在一个路由之上6LoWPAN间实施DAD，这里重用上面的地址注册选项。

7）在一个路由之上网络间分发前缀和语境信息的新的可选机制，这里使用一个新的权威边界路由器选项来控制配置改变的洪泛（过程）。

针对LoWPAN的这个IPv6 ND优化，目前在IETF是一项正在进行的工作。

8.7.3 6LoWPAN中的IPv6地址自动配置

自动配置6LoWPAN中的一个IPv6地址，方法是通过连接一个64比特接口标识符，该标识符从制造商指派到IEEE 802.15.4设备的一个EUI-64标识符派生得到，或由PAN协调器指派到64比特网络前缀（链路本地或全局的）的16比特短地址派生得到。依据“以太网上的IPv6”规范[4]，从EUI-64形成接口标识符。所有IEEE 802.15.4设备有一个EUI-64地址，但16比特短地址也是可能的。在这些情形中，形成一个“伪48比特地址”，如图8.12所示。

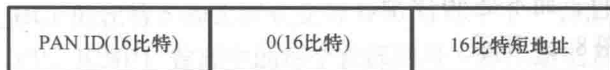


图8.12 由16比特短地址形成48比特地址

依据“以太网之上的IPv6”规范[4]，由这个48比特地址形成接口标识符。但是，在得到的接口标识符中，“全局/本地”（U/L）比特必须被设置为0，以便保持这样的事实，即这不是一个全局唯一的值。对于任一地址格式，一定不要使用全零地址。

8.7.4　首部压缩

典型情况下，为点到点链路场景定义了首部压缩技术，其中压缩器和解压缩器是相互直接地和排他地进行通信。但是，人们期望，IEEE 802.15.4设备将被部署在多跳网络中。所以，人们高度期望的是，IEEE 802.15.4网络中的一台设备能够通过它的任意邻居发送首部压缩的分组，在尽可能少的初步语境构造下做到这一点。首部压缩可能导致没有落在一个字节边界的对齐问题。因为典型情况下，硬件不能以传输小于一个字节的单元来传输数据，所以必须使用填充。

在一个给定6LoWPAN中所有节点共享与IPv6链路（IEEE 802.15.4 PAN）有关的相同信息这个事实基础上，6LoWPAN为首部压缩定义了两种方法。

无状态首部压缩：为在6LoWPAN内使用的链路本地IPv6通信，优化了这种方法。这种方法定义HC1压缩IPv6首部，定义HC2压缩上层首部。这种方法不压缩跳限制、全局IPv6地址和组播地址。

有状态的或基于语境的首部压缩：这种压缩方法为压缩IPv6首部定义了一种较好的压缩方案，称作IPv6首部压缩（IPHC）。

1. 在无状态首部压缩（HC1）中IPv6首部字段的编码

HC1首部压缩主要适用于链路本地通信。这种方法不保持任何流状态。事实上，它依赖于与整条链路有关的信息。如下IPv6首部字段是已知的，或它们可从链路层信息推断得到。

1）IPv6版本v6。

2）IPv6源和目的地址是链路本地的，源或目的地址的IPv6接口标识符（低64比特）可从层2源和目的地址推断得到。

3）分组长度可从层2帧长度或从分片首部中的“数据报尺寸”（如果存在的话）推断得到。

4）流量类和流标签都为0。

5）下一首部为UDP、因特网控制消息协议（ICMP）或TCP。

在IPv6首部中总是需要完整携带的唯一字段是跳限制（8比特）。

链路本地通信的这个IPv6首部可被压缩到2字节（1字节用于HC1编码，1字节用于跳限制），而不是40字节。

HC1编码如图8.13所示。

由HC1编码进行编码的地址字段解释如下。

1）IPv6源地址（比特0和比特1）：

不受影响的无线电球形空间内任何其他网络所用的一个 PAN 标识符，做到这一点。一旦选中 PAN 标识符，PAN 协调器支持其他设备（潜在地有 FFD 和 RFD）加入它的网络。星形拓扑的典型应用有家庭自动化、个人计算机（PC）外设、玩具和游戏以及个人保健。

2. 对等网络拓扑

对等拓扑也有一个 PAN 协调器。但是，它不同于星形拓扑的是，任意设备可与任何其他设备通信，条件是只要它们在相互的通信范围内。对等拓扑支持实现更复杂的网络形成，如网状联网拓扑。诸如工业控制和监测、无线传感器网络、资产和库存跟踪、智能农业以及安全等应用，将受益于这样一种网络拓扑。一个对等网络可以是独特的、自组织的和自愈的。它也支持多跳将消息从任意设备路由到网络上的任何其他设备，其中使用网状路由技术，这典型地在高层协议的帮助下做到这一点。

对等通信拓扑的一个例子是集群树，如图 8.5 所示。集群树网络是对等网络的一个特殊情形，其中多数设备是 FFD。一个 RFD 作为一个分支末端的一个叶设备连接到一个集群树网络，原因是 RFD 不支持关联其他设备。任何 FFD 可作为一个协调器，并提供到其他设备的同步服务。PAN 协调器形成第一个集群，方法是选择一个未用 PAN 标识符，并将信标帧广播到邻居设备。

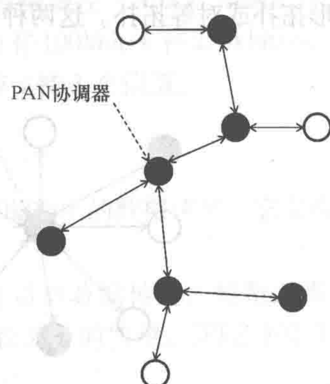


图 8.5 集群树拓扑

接收到一个信标帧的一个候选设备可在 PAN 协调器处请求加入网络。如果 PAN 协调器允许该设备加入，则它将新设备作为一个子设备加入到它的邻居列表。之后新加入的设备将 PAN 协调器作为父设备加入到它的邻居列表，并开始传输周期性的信标，之后其他后续设备可在那台设备处加入网络。

8.6 IPv6

将在 IEEE 802.15 MAC 数据帧上携带 IPv6 分组，典型情况下，建议 IPv6 分组在请求确认的帧中加以携带。一个 IEEE 802.15.4 PAN 被处理为 IPv6 操作的单一 IPv6 链路。

在 IEEE 802.15.4 上传输 IPv6 分组的 MTU 尺寸是 1280 字节，这是最小 IPv6 MTU。但是，一个完整的 IPv6 分组不能装在一个 IEEE 802.15.4 帧中。IEEE 802.15.4 协议数据单元有不同尺寸，取决于存在多少开销。

最大物理层分组尺寸是 127 字节。

IEEE 802.15.4 中的最大帧是 25 字节。

最大链路层安全（AES-CCM-128 情形）开销是 21 字节。

可用于 IPv6 分组的净荷尺寸是 81（127 - 25 - 21）字节。

考虑一个 40 字节 IPv6 首部和一个 8 字节 UDP 首部，可用于应用的实际净荷尺寸是 33 字节。

上述考虑使 IPv6 在其原始形式中运行在 IEEE 802.15.4 之上是非常低效的，并导致如下观察结果：

1）必须提供适配层，满足 IPv6 对一个最小 MTU 的需求。但是，人们期望 IEEE 802.15.4 的多数应用将不适用这种大型的分组，且小型应用净荷与合适的首部压缩一起使用，将产生适合在单个 IEEE 802.15.4 帧内的分组。

2）即使上述空间计算给出最坏场景，但它确实指出这样的事实，即对于 IPv6 在 IEEE 802.15.4 之上传输，首部压缩是紧迫要求，是不可避免的。

8.7　6LoWPAN：在无线个域网之上传输 IPv6

一个 6LoWPAN 定义在 IEEE 802.15.4 之上传输 IPv6，特别定义了 IPv6 分组传输的帧格式，以及 IPv6 链路本地和全局地址的形成。因为 IPv6 要求比 IEEE 802.15.4 帧尺寸大得多的分组尺寸，为分片和重新组装定义了一个适配层。一个 6LoWPAN 也为首部压缩定义了机制，这是在 IEEE 802.15.4 网络之上实际传输 IPv6 所需的，也定义了在 IEEE 802.15.4 网状网中分组交付所需的就绪提供机制。

8.7.1　LoWPAN 帧格式和交付

6LoWPAN 为在 IEEE 802.15.4 帧中高效封装 IPv6 数据报定义了各种首部。在 IEEE 802.15.4 之上传输的所有 6LoWPAN 封装的数据报，外部加上一个封装首部栈。首部栈每个首部包含一个首部类型，后跟零个或多个首部字段。6LoWPAN 首部的类型如图 8.6 所示。

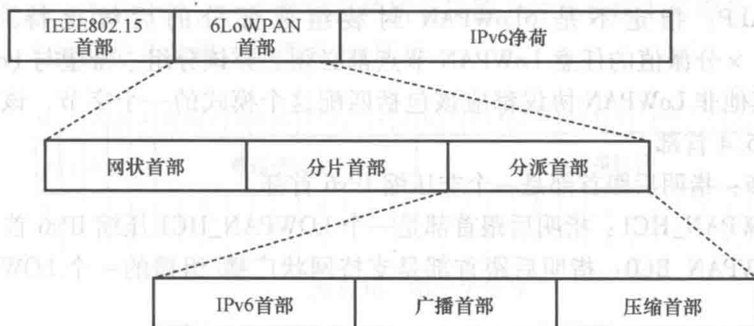


图 8.6　6LoWPAN 首部

不是所有首部在所有时间都存在。如果存在，它们必须按如下顺序出现：

1）网状寻址首部。

2）广播分派首部。

3）分片首部。

4）压缩（HC1）首部。

所有协议数据报（IPv6，压缩IPv6首部等）必须在前面有有效6LoWPAN封装首部之一。

1. 6LoWPAN分派首部

分派首部的格式如图8.7所示。

分派类型（2比特）	分派（6比特）	分派特定的首部

图8.7　6LoWPAN分派首部

分派首部的字段定义如下：

1）分派类型：0比特被定义为第一比特，1比特被定义为第二比特。

2）分派：是一个6比特选择器，识别直接后跟分派首部的首部类型。

3）分派特定的首部：是由分派首部确定的一个首部。

图8.8给出了为一个6LoWPAN定义的分派首部列表。

分派首部（前8比特）	描述
00 xxxxxx	NALP：不是一个LoWPAN帧
01 000001	IPv6：非压缩IPv6地址
01 000010	LOWPAN_HC1：LOWPAN_HC1压缩的IPv6
01 010000	LOWPAN_BC0：LOWPAN_BC0广播
01 111111	ESC:后跟其他分派字节
所有其他值	保留的：为未来用途保留

图8.8　分派值比特模式

1）NALP：指定不是6LoWPAN封装组成部分的后续比特，遇到00××××××分派值的任意LoWPAN节点都必须丢弃该分组。希望与LoWPAN节点共存的其他非LoWPAN协议都应该包括匹配这个模式的一个字节，该字节直接后跟802.15.4首部。

2）IPv6：指明后跟首部是一个非压缩IPv6首部。

3）LOWPAN_HC1：指明后跟首部是一个LOWPAN_HC1压缩IPv6首部。

4）LOWPAN_BC0：指明后跟首部是支持网状广播/组播的一个LOWPAN_BC0首部。

5）ESC：指明后跟首部是分派值的单个8比特字段。它支持大于127的分派值。

2. 网状寻址类型和首部

网状类型由1和0定义为前两个比特。网状类型首部如图8.9所示。



图8.9　网状寻址类型和首部

字段定义如下：

1）V：一个1比特字段。如果源发［或"恰好是第一个"（very first）］地址是一个IEEE EUI-64地址，则为0；如果它是一个短16比特地址，则为1。

2）F：一个1比特字段。如果最终目的地址是一个IEEE EUI-64地址，则为0；如果它是一个短16比特地址，则为1。

3）剩下的跳数：一个4比特字段。在将这条分组往下一跳发送之前，每个转发节点将这个值减1。如果剩下的跳数减少到0，则分组不再进一步转发。值0×F被保留，并指明直接后跟的一个8比特剩余的深度跳字段，支持一个源节点指明大于14跳的一个跳限制。

4）源发者地址：源发者的链路层地址。

5）最终目的地址：最终目的地的链路层地址。

注意，V和F比特支持16比特和64比特MAC地址构成的一个混合地址。这是有用的，至少支持网状层"广播"，原因是802.15.4广播地址被定义为16比特短地址。

3. 分片首部

如果整个净荷（IPv6数据报）可封装在单个802.15.4帧内，则它不分片，且LoWPAN封装不应包含一个分片首部。如果数据报不能封装在单个IEEE 802.15.4帧内，则它必须被分解为多个链路分片。分片偏移是以8字节的整数倍表示的，所以除了最后一个分片外，一个数据报的所有链路分片在长度上必须是8字节的整数倍。第一个链路分片的格式如图8.10所示。



图8.10　第一个分片

第二个和后续的链路分片（直到最后一个分片并包括该分片）必须包含一个

分片首部，它符合如图 8.11 所示的格式。

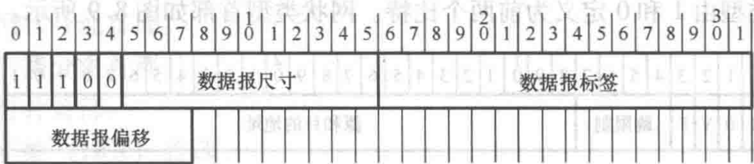


图 8.11 第二个和后续分片

分片首部字段定义如下：

1）数据报尺寸：这个 11 比特字段对链路层（6LoWPAN）分片之前（但在 IP 层分片之前，如果有的话）的整个 IP 分组尺寸进行编码。对于一个 IPv6 分组的所有链路层分片，数据报尺寸的值必须是一样的。

2）数据报标签：这个字段是 16 比特长，且初始值是没有确定的，一台设备可检取一个随机初始值。对于一个净荷（如 IPv6）数据报的所有链路分片，数据报标签的值必须是相同的。对于后续的被分片数据报，发送方必须增加数据报标签的值。数据报标签被增加的值必须从 65535 环回到 0。

3）数据报偏移：这个字段仅存在于第二个和后续链路分片中，必须指明偏移，是以 8 字节的增量表示的，指从净荷数据报的开始部分的分片偏移。这个字段有 8 比特长。

链路分片的接收方将使用如下信息，重构 IPv6 数据报：

1）发送方的 802.15.4 源地址（或在一个网状配置中源发者的地址）。

2）目的地的 802.15.4 地址（或在一个网状配置中的最终目的地地址）。

3）数据报尺寸。

4）数据报标签，识别属于一个给定数据报的所有链路分片。

在接收到一个链路分片时，接收方开始构造原始未分片分组，其尺寸是数据报尺寸。它使用数据报偏移字段确定在原始未分片分组内各分片的位置。当一个节点第一次接收到带有一个给定数据报标签的一个分片时，它启动一个重组定时器。当这个定时器超时时，如果整个分组没有被重新组装，则丢弃已存在的分片。重新组装超时被设置为 60s 的最大值（这也是 IPv6 重新组装规程[3] 中的超时）。

8.7.2 一个 6LoWPAN 的邻居发现

IPv6 邻居发现（ND）[6] 是 IPv6 主机发现在线（on-link）路由器、前缀和其他配置信息的一个信令协议。它也为地址解析和重复地址检测（DAD）指定机制。

因为 LoWPAN 无线电链路是丢失性的，展示出间歇性连接，所以 6LoWPAN 的 IPv6 ND 必须就如下方面做出优化：

1）通过避免使用组播洪泛并降低使用链路范围组播消息，以此最小化信令。

2）优化主机及其缺省路由器之间的接口。

3）支持正处于睡眠的主机。

4）最小化节点的复杂性。

5）依据首部压缩技术所需，将语境信息传播到各主机。

一个6LoWPAN的ND优化适用于网状网之下（mesh-under）和路由之上（route-over）配置。在网状网下的配置中，仅存在6LoWPAN边界路由器和主机；在网状网下拓扑中，不存在6LoWPAN路由器。

优化的最重要部分是演进的主机到路由器交互通信，这支持睡眠中的节点，并避免使用组播ND消息，例外情况是一台主机找到缺省路由器的一个初始集合并重新做出这种决定的情形，此时那些路由器集已经不可达。

对IPv6邻居发现的扩展（RFC 4861）

6LoWPAN ND指定对IPv6 ND的如下优化和扩展[6]。

1）路由器通告信息的主机初始刷新。这去除了从主机到主机的周期性的或非请求路由器通告的需求。

2）如果使用基于EUI-64的IPv6地址，则不需要DAD。

3）如果使用DHCPv6[17]来指派地址，则DAD是可选的。

4）一种使用主机和路由器之间新的地址注册选项的新地址注册机制。这消除了路由器使用组播邻居请求来发现主机并支持睡眠中主机的需要。这也支持相同的IPv6地址前缀用在一个路由之上6LoWPAN间。它为DAD提供主机到路由器接口。

5）一个新的可选路由器通告选项，用于6LoWPAN首部压缩所用的语境。

6）一个新的可选机制，在一个路由之上6LoWPAN间实施DAD，这里重用上面的地址注册选项。

7）在一个路由之上网络间分发前缀和语境信息的新的可选机制，这里使用一个新的权威边界路由器选项来控制配置改变的洪泛（过程）。

针对LoWPAN的这个IPv6 ND优化，目前在IETF是一项正在进行的工作。

8.7.3 6LoWPAN中的IPv6地址自动配置

自动配置6LoWPAN中的一个IPv6地址，方法是通过连接一个64比特接口标识符，该标识符从制造商指派到IEEE 802.15.4设备的一个EUI-64标识符派生得到，或由PAN协调器指派到64比特网络前缀（链路本地或全局的）的16比特短地址派生得到。依据“以太网上的IPv6”规范[4]，从EUI-64形成接口标识符。所有IEEE 802.15.4设备有一个EUI-64地址，但16比特短地址也是可能的。在这些情形中，形成一个“伪48比特地址”，如图8.12所示。

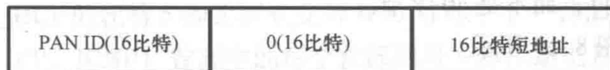


图8.12 由16比特短地址形成48比特地址

依据“以太网之上的IPv6”规范[4]，由这个48比特地址形成接口标识符。但是，在得到的接口标识符中，“全局/本地”（U/L）比特必须被设置为0，以便保持这样的事实，即这不是一个全局唯一的值。对于任一地址格式，一定不要使用全零地址。

8.7.4 首部压缩

典型情况下，为点到点链路场景定义了首部压缩技术，其中压缩器和解压缩器是相互直接地和排他地进行通信。但是，人们期望，IEEE 802.15.4设备将被部署在多跳网络中。所以，人们高度期望的是，IEEE 802.15.4网络中的一台设备能够通过它的任意邻居发送首部压缩的分组，在尽可能少的初步语境构造下做到这一点。首部压缩可能导致没有落在一个字节边界的对齐问题。因为典型情况下，硬件不能以传输小于一个字节的单元来传输数据，所以必须使用填充。

在一个给定6LoWPAN中所有节点共享与IPv6链路（IEEE 802.15.4 PAN）有关的相同信息这个事实基础上，6LoWPAN为首部压缩定义了两种方法。

无状态首部压缩：为在6LoWPAN内使用的链路本地IPv6通信，优化了这种方法。这种方法定义HC1压缩IPv6首部，定义HC2压缩上层首部。这种方法不压缩跳限制、全局IPv6地址和组播地址。

有状态的或基于语境的首部压缩：这种压缩方法为压缩IPv6首部定义了一种较好的压缩方案，称作IPv6首部压缩（IPHC）。

1. 在无状态首部压缩（HC1）中IPv6首部字段的编码

HC1首部压缩主要适用于链路本地通信。这种方法不保持任何流状态。事实上，它依赖于与整条链路有关的信息。如下IPv6首部字段是已知的，或它们可从链路层信息推断得到。

1）IPv6版本v6。

2）IPv6源和目的地址是链路本地的，源或目的地址的IPv6接口标识符（低64比特）可从层2源和目的地址推断得到。

3）分组长度可从层2帧长度或从分片首部中的“数据报尺寸”（如果存在的话）推断得到。

4）流量类和流标签都为0。

5）下一首部为UDP、因特网控制消息协议（ICMP）或TCP。

在IPv6首部中总是需要完整携带的唯一字段是跳限制（8比特）。

链路本地通信的这个IPv6首部可被压缩到2字节（1字节用于HC1编码，1字节用于跳限制），而不是40字节。

HC1编码如图8.13所示。

由HC1编码进行编码的地址字段解释如下。

1）IPv6源地址（比特0和比特1）：

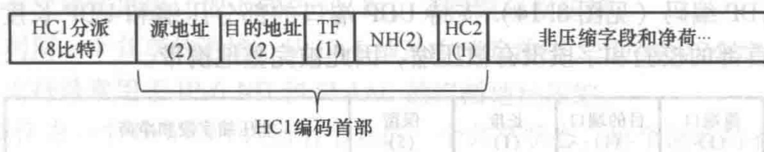


图 8.13 LOWPAN_ HC1 编码

00：PI、II

01：PI、IC

10：PC、II

11：PC、IC

2）IPv6 目的地址（比特 2 和比特 3）：

00：PI、II

01：PI、IC

10：PC、II

11：PC、IC

图例：

PI：在没有压缩条件下在线路中携带的前缀。

PC：前缀压缩，在分组中假定链路-本地前缀，并加以消除。

II：在线路中携带的接口标识符。

IC：消除的接口标识符，从相应的链路层地址中可推断得到。

3）流量类和流标签（TF：比特 4）：

0：没有压缩，对流量类有完整的 8 比特，对要发送的流标签有 20 比特。

1：流量类和流标签都为 0。

4）下一首部（NH：比特 5 和比特 6）：

00：没有压缩，要发送完整的 8 比特。

01：UDP

10：ICMP

11：TCP

5）HC2 编码（HC2：比特 7）：

0：没有更多的首部压缩比特。

1：依据 HC2 编码格式，HC1 编码直接后跟更多的首部压缩比特。

2. 无状态首部压缩（HC2）中 UDP 首部字段的编码

LOWPAN_HC1 的比特 5 和比特 6 支持 IPv6 首部中下一首部字段的压缩。这些协议（UDP、TCP、ICMP）首部中的每个首部的进一步压缩也是可能的。下面介绍 UDP 首部本身是如何被压缩的。本节中的 HC2 编码是 HC_UDP 编码，它仅适用于这样的情况，即 HC1 中的比特 5 和比特 6 指明后跟 IPv6 首部的协议是 UDP。

HC_UDP 编码（见图8.14）支持 UDP 端口的部分压缩和 UDP 长度的完整压缩。UDP 首部的校验和字段没有被压缩，因此被完整地携带。



图8.14　使用HC2的UDP首部编码

源和目的端口的部分压缩是在端口范围（61616～61631）的基础上得到的，且UDP长度信息可如下推断得到，即IPv6首部的净荷长度字段减去IPv6首部和UDP首部之间存在的任何扩展首部的长度。这种方法支持将UDP首部压缩到4字节而不是原始的8字节。

1）UDP 源端口（比特0）：

0：不压缩，在线路中携带。

1：实际的16比特源端口是通过计算P+短端口值得到的。P的值是数61616（0xF0B0）。短端口表示为在线路中携带的一个4比特值。

2）UDP 目的端口（比特1）：

0：没有压缩，在线路中携带。

1：压缩到4比特，实际的16比特目的端口是通过计算P+短端口值得到的。P的值是数61616（0xF0B0）。短端口表示为在线路中携带的一个4比特值。

3）长度（比特2）：

0：没有压缩，在线路中携带。

1：压缩的，从IPv6首部长度信息中计算得到长度。

4）预留的（比特3到比特7）：

这5个比特是为未来用途预留的。

3. 有状态的或基于语境的首部压缩

作为基于语境的首部压缩的组成部分，定义了两种编码格式：LoWPAN IPHC 和 LoWPAN 下一首部压缩（NHC）。IPHC 被用来压缩IPv6首部，而NHC被用来压缩后跟IPv6首部的任意下一首部。

IPHC定义一个改进的编码格式（见图8.15），用来压缩IPv6首部，并提供在前面定义的无状态首部压缩之上的如下改进：

IPHC类型 (3)	TF (2)	NH (1)	HLIM (2)	CID (1)	SAC (1)	SAM (2)	M (1)	DAC (1)	DAM (2)

图8.15　6LoWPAN改进的IPv6首部压缩

1）支持流量类和流标签字段，可独立地进行压缩。

2）当使用常见值（如1或255）时，支持跳限制压缩。

3）利用一个共享的语境，从 IPv6 地址中消除前缀，其中包括全局 IPv6 地址。

4）支持最常用于 IPv6 ND 和 SLAAC 的组播地址压缩。

语境作为一个 LoWPAN 内所有节点的一个共享状态。单个语境持有单个前缀。IPHC 使用一个 4 比特索引识别一个语境，使 IPC 能够在一个 LoWPAN 内同时支持高达 16 个语境。当一个 IPv6 地址匹配一个语境的存储前缀时，IPHC 将前缀压缩到语境的 4 比特标识符。可为任意前缀配置共享的语境，从而使 LoWPAN 中的各节点可压缩源地址和目的地址中的前缀，即使当与 LoWPAN 外的节点通信时也是如此。

1）IPHC 类型：3 比特字段（011），指明 IPHC 首部类型。

2）TF：流量类、流标签：

00：显式的拥塞通知（ECN）+区分服务码点（DSCP）+4 比特填充 + 在线路中携带的流标签（4 字节）。

01：ECN +2 比特填充 + 在线路中携带的流标签（3 字节）；消除 DSCP。

10：ECN + 在线路中携带的 DSCP（1 字节）；消除流标签。

11：消除流量类和流标签。

3）NH：下一首部：

0：下一首部的完整 8 比特，是在线路中携带的。

1：压缩下一首部字段，使用 LoWPAN NHC 对下一首部编码，这在下一部分“4. UDP 的 LoWPAN NHC 编码”中讨论。

4）HLIM：跳限制：

00：跳限制字段是在线路中携带的。

01：压缩跳限制字段，跳限制是 1。

10：压缩跳限制字段，跳限制是 64。

11：压缩跳限制字段，跳限制是 255。

5）CID：语境标识符压缩：

0：不使用额外的 8 比特 CID 扩展。如果在源地址压缩（SAC）或目的地址压缩（DAC）中指明基于语境的压缩，则使用语境 0。

1：一个额外的 8 比特语境标识符扩展字段，直接后跟目的地址模式（DAM）字段。

6）SAC：源地址压缩：

0：SAC 使用无状态压缩。

1：SAC 使用有状态的、基于语境的压缩。

7）SAM：源地址模式：

① 如果 SAC =0：

00：128 比特。完整地址是在线路中携带的。

01：64 比特。消除地址的前 64 比特。那些比特的值是链路本地前缀，以 0 填充。其他 64 比特是在线路中携带的。

10：16 比特。消除地址的前 112 比特。那些比特的值是链路本地前缀，以 0 填充。其他 16 比特是在线路中携带的。

11：0 比特。地址被完全消除。地址的前 64 比特是以零填充的链路本地前缀。剩下的 64 比特是从链路层地址计算得到的。

② 如果 SAC = 1：

00：一个未指派的地址。

01：64 比特。使用语境信息和在线路中携带的 64 比特，推导得到地址。

10：16 比特。使用语境信息和在线路中携带的 16 比特，推导得到地址。

11：0 比特。完全消除地址。前缀是使用语境信息推导得到的。不能由语境信息涵盖的其他剩余 64 比特是从链路层地址计算得到的。

8）M：组播压缩：

0：目的地址不是一个组播地址。

1：目的地址是一个组播地址。

9）DAC：目的地址压缩：

0：DAC 使用无状态压缩。

1：DAC 使用有状态、基于语境的压缩。

10）DAM：目的地址模式：

① 如果 M = 0 且 DAC = 0，则这种情形匹配 SAC = 0，但对于目的地址

00：128 比特。完整的地址是在线路中携带的。

01：64 比特。消除地址的前 64 比特。那些比特的值是以零填充的链路本地前缀。剩余的 64 比特在线路中携带。

10：16 比特。消除地址的前 112 比特。那些比特的值是以零填充的链路本地前缀。剩余的 16 比特在线路中携带。

11：0 比特。完全消除地址。地址的前 64 比特是以零填充的链路本地前缀。剩余的 64 比特是从链路层地址计算得到的。

② 如果 M = 0 且 DAC = 1：

00：预留的。

01：64 比特。使用语境信息和在线路中携带的 64 比特，推导得到地址。

10：16 比特。使用语境信息和在线路中携带的 16 比特，推导得到地址。

11：0 比特。完全消除地址。使用语境信息推导得到前缀。没有由语境信息涵盖的其他剩余 64 比特，是从链路层地址计算得到的。

③ 如果 M = 1 且 DAC = 0：

00：128 比特。完整的地址是在线路中携带的。

01：48 比特。地址采取 FF × ×::00 × ×：× × × ×：× × × ×的形式。

10：32 比特。地址采取 FF××::00××:××××的形式。

11：8 比特。地址采取 FF02::00××的形式。

④ 如果 M=1 且 DAC=1：

00：48 比特。这个格式被设计来匹配基于单播前缀的 IPv6 组播地址，见参考文献［20，21］中的定义。组播地址采取形式 FF××:××LL:PPPP:PPPP:PPPP:PPPP:××××:××××，其中×是在线路中携带的半字节，是以它们在这个格式中的顺序出现的。P 表示用来编码前缀本身的各个半字节。L 表示用来对前缀长度编码的各个半字节。前缀信息 P 和 L 是从指定的语境中取出的。

01：保留的。

10：保留的。

11：保留的。

4. UDP 的 LoWPAN NHC 编码

像在 HC2 中一样，NHC 利用相同的端口范围（61616~61631），在最佳情形中，高效地将每个 UDP 端口压缩到 4 比特，并去除 UDP 净荷长度字段，原因是它总是可使用 6LoWPAN 分片首部或 IEEE 802.15.4 首部从低层推导得到。NHC 也支持消除 UDP 校验和，此时一个高层消息完整性校验涵盖相同的信息，至少具有相同的长度。当使用传输或应用层安全时，这样一个场景是典型情况。结果，在最佳情形中，UDP 首部可被压缩到 2 字节。UDP 首部的 NHC 如图 8.16 所示。



图 8.16 UDP 的 NHC 编码首部

1）C：校验和：

0：校验和的所有 16 比特都是在线携带的。

1：去除校验和的所有 16 比特。通过在 6LoWPAN 终结点上重新计算校验和而恢复它。

2）P：端口：

00：用于源端口和目的端口的所有 16 比特都是在线路中携带的。

01：用于源端口的所有 16 比特是在线路中携带的。目的端口的前 8 比特是 0×F0并被消除。目的端口的剩余 8 比特是在线路中携带的。

10：源端口的前 8 比特是 0×F0 并被消除。源端口的剩余 8 比特是在线路中携带的。目的端口的所有 16 比特都是在线路中携带的。

11：源端口和目的端口的前 12 比特是 0×F0B，并被消除。每个剩下的 4 比特都是在线路中携带的。

在线路中携带的各字段（部分的或整体上的）以它们出现在 UDP 首部格式中

的相同顺序出现[9]。UDP 长度字段必须总被消除，可使用 6LoWPAN 分片首部或 IEEE 802.15.4 首部从低层推断得到。

8.7.5 6LoWPAN 网状路由

即使人们期望 IEEE 802.15.4 网络为一些应用使用网状路由，IEEE 802.15.4 规范也没有定义这样的能力。为在 LoWPAN 中取得 IPv6 分组的多跳传输，6LoWPAN 已经定义了一种链路层机制。在一个 6LoWPAN 中，可实施两种类型的网状路由，即网状网之下（mesh-under）和网状网之上（mesh-over）。网状网之下路由是层 2 转发，并使用网状首部进行帧交付，而网状网之上路由是使用 IPv6 的 IP 层路由。

在网状网之下路由（见图 8.17）中，通过在 LoWPAN 封装的任何其他首部之前包括一个网状寻址首部，支持网状网交付。分组源发方将网状网寻址首部中的源发方链路层地址设置为自己的链路层地址，将最终目的地的链路层地址设置为分组的最终目的地。它将 802.15.4 首部中的源地址设置为自己的链路层地址，并将转发者的（在下一跳的 LoWPAN 节点）链路层地址放入 802.15.4 首部的目的地址字段。最终，它传输该分组。

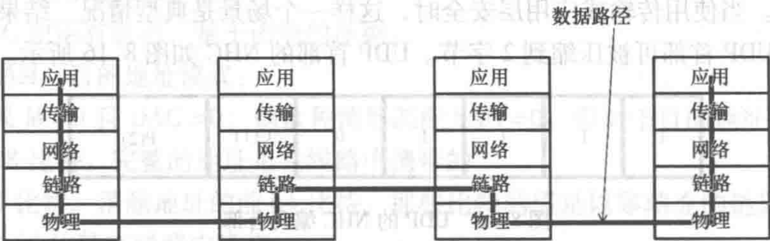


图 8.17 6LoWPAN 网状网之下路由

类似地，如果一个节点接收到带有一个网状寻址首部的帧，则它必须查看网状寻址首部的最终目的地字段，确定真实的目的地。如果节点本身是最终目的地，则它依据正常的交付而消耗该分组。如果它不是最终目的地，则设备减少剩余跳字段，如果结果为零，则它丢弃该分组。否则，该节点咨询其链路层路由表，确定去往最终目的地的下一跳应该是什么，并将那个地址放入 802.15.4 首部的目的地址字段。最后，该节点将 802.15.4 首部中的源地址改变为自己的链路层地址，并传输该分组。

与网状网之下路由不同的是，网状网之上路由（见图 8.18）是使用 IP 路由在网络层实施的。当一个 LoWPAN 已经使用不同的链路层技术构造时，这种类型的网状网是有用的。

人们仅期望 FFD 参与作为一个网状网中的路由器。FFD 将它们自己限制为发现 FFD，并使用这些 FFD 实施它们所有的转发，这是以类似于 IP 主机如何典型地



图8.18 6LoWPAN 网状网之上路由

使用默认路由器转发它们的所有离线（off-link）流量的方式进行的。对于使用网状网交付的一个RFD，“转发器”总是合适的FFD。

8.7.6 LoWPAN 广播

在未来使用6LoWPAN中的广播机制，可能得到另外的功能。使用一个广播首部，做到一个6LoWPAN中的广播，如图8.19所示，它由一个LOWPAN_BC0分派，后跟一个序列号字段组成。序列号被用来检测重复分组，并抑制重复分组。



0	1	LOWPAN_BC0 (6比特)	序列号 (8比特)

图8.19 广播首部

字段定义如下：

1）LOWPAN_BC0：6比特（010000）。

2）序列号：一个8比特字段，当源发方发送一条新的网状网广播或组播分组时，它将该字段做加1处理。

8.8 传输层

TCP是最广泛用于万维网的传输协议。但是，由TCP使用的流控机制对变化的时延是非常敏感的，不能良好地适用于6LoWPAN，6LoWPAN典型地由丢失性和不确定的链路组成。为降低TCP开销，在6LoWPAN中也许首选UDP。但是，UDP不能保障分组交付，它要求应用层支持。为支持在丢失性的、网状网路由的链路上的传输，针对增强TCP的研究工作也在进行，这支持在各种媒介上服务的无缝部署，这些媒介并不总是像以太网链路那样动作。

LoWPAN的新应用协议（见8.9节），称作受约束的应用协议（CoAP），是由IETF开发的，默认地假定UDP作为传输协议，也可选地运行在TCP之上。

8.9 应用层协议

在因特网上使用web服务，在多数应用中已经成为泛在情形，为LoWPAN使用IPv6的基本思路是让LoWPAN中的设备与其他IP连接的节点直接交互通信。如今，因特网上的多数IP连接的节点，支持RESTful架构，该架构基于TCP（用于数据的可靠传递）之上的超文本传输协议（HTTP）（用于资源的检索和操作），并使用标准的基于文本的消息格式［像可扩展标记语言（XML）或超文本标记语言（HTML）］来构造数据。在企业网中，可使用简单对象访问协议（SOAP）而不是RESTful架构。在基于IP网络中经常使用的其他协议有服务定位协议（SLP）[18]和简单网络管理协议（SNMP）[19]。SLP提供一个灵活的和可扩展的框架，为主机提供对联网服务的存在、位置和配置的信息，特别在企业网中更是如此。

但是，RESTful架构，特别是HTTP，是基于请求/响应范型的，不适合LoWPAN中的资源受限设备（如带有有限RAM和ROM的8/16比特微处理器），其中节点的占空周期为0.1%或更低。另外，HTTP分组尺寸为LoWPAN中可能的通常50~60字节净荷，施加额外的挑战。

通过使用万维网联盟（W3C）高效XML互换（EXI）编码，LoWPAN设备中的分组尺寸限制可一定程度地加以克服。但是，IEEE 802.15.4设备之上的HTTP和TCP性能带来挑战。这使IETF定义称作CoAP的一个新协议。

CoAP类似于HTTP，并基于RESTful架构，以便用于像LoWPAN的受约束网络。CoAP提供到HTTP的方便转换，以便与万维网集成，同时满足专门化的需求，像组播支持、非常低的开销和针对受约束环境的简单性。CoAP有如下主要特征：

1）基于RESTful架构的一项设计，这最小化与HTTP的映射复杂度。

2）低首部开销和剖析复杂度。

3）统一资源标识符（URI）和内容类型支持。

4）对资源发现的支持。

5）对一项资源的简单订阅，并产生推送通知。

6）基于最大年龄的简单缓存。

也定义了采用HTTP映射CoAP，这支持构建代理，以一种统一的方式通过HTTP提供到CoAP资源的访问。CoAP默认地工作在UDP之上，并可选地工作在TCP之上，以便传输大量数据。

CoAP支持CREATE、UPDATE、READ和DELETE（CRUD）的基本RESTful方法，这些方法可容易地映射到HTTP方法。另外，引入了称作NOTIFY（通知）的一个推送方法，发布时间并报告其他信息。CoAP方法操控资

源，它们具有安全（只能检索）和幂等（意味着多个同样的请求应该与单条请求具有相同效果）的性质。READ 方法是安全的。因此，除了检索外，它对一项资源不会采取任何其他动作。READ、UPDATE、DELETE 和 NOTIFY 方法可被看作是幂等的。

当 CoAP 运行在 UDP 之上时，整条消息可装在单条数据报内。当与 6LoWPAN 一起使用时，消息装到单条 IEEE 802.15.4 帧中，以便最小化分片。

8.10 物联网的网络架构

该架构由各种传感器/中继/执行器（为控制和管理目的而产生/消耗信息）、参与到路由中的网状网节点以及网关/边缘/边界路由器（汇聚流量，并连接到一台或多台管理服务器及因特网）组成。典型情况下，LoWPAN 可工作在三种类型的连接模型中，这取决于应用和需要。它们有：

1）自治 LoWPAN。

2）带有扩展因特网连接能力的 LoWPAN。

3）真正的物联网。

8.10.1 自治 LoWPAN

并不总是要求一个 LoWPAN 连接到因特网。例如，如图 8.20 所示，一个工厂监测系统仅连接到一台本地管理服务器，并选择不连接到因特网，原因是不需要与外部世界通信。



图 8.20 自治地工作的一个 LoWPAN

8.10.2 具有扩展因特网连接能力的 LoWPAN

通过使用一台防火墙和代理服务器，一些 LoWPAN 可提供到 LoWPAN 设备的有限的和受控的访问，如图 8.21 所示。这样的扩展连接能力，对于通知事件和告警以及检索信息和来自一个远端位置的控制，是有用的。



图8.21 带有到外部世界（在因特网之上）受控访问的一个LoWPAN

8.10.3 真正的物联网

在这个模型中，所有设备/物体在因特网之上都是可见的，可与因特网上的其他设备直接采用端到端安全进行通信。如图8.22所示，一台边缘路由器提供HTTP和CoAP之间的转换。

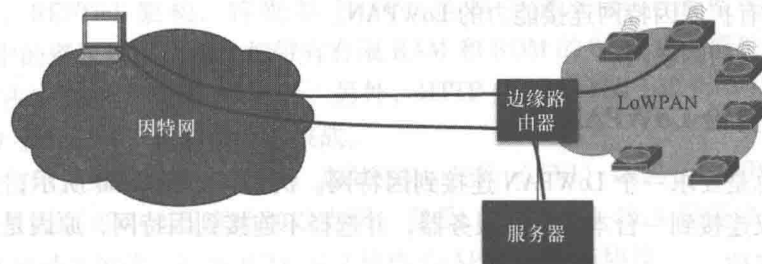


图8.22 一个真正的物联网

8.11 安全考虑

像任何其他无线网络一样，LoWPAN对被动的窃听攻击和可能甚至是主动的篡改，是脆弱的，原因是为参与到通信的过程，是不要求物理接入到导线的。设备及其成本目标（有限的CPU、RAM、ROM和电池）之间的短期关系，施加额外的安全约束，这使这些网络成为要保障安全的最困难环境。这些约束也许严重地限制密码学算法和协议的选择，并将影响安全架构的设计。另外，电池寿命和成本约束对这些网络可容忍的安全开销，施加严重的限制。

安全是物联网的一项重要需求。虽然IETF为保障IP网络的安全而定义了几种方法，如因特网协议安全（IPsec）、传输层安全（TLS）等，但这些安全协议如何高效地集成到资源受限的LoWPAN中，对于物联网是至关重要的。当前可用的安全方法有：

1）在 IEEE 802.15.4 链路层处的 AES-128 加密。

2）在 IP 层的 IPsec。

3）在应用层的 TLS。

但是，这些机制中的哪些机制采用非常短寿命的消息序列可成功地应用到资源受限的设备和网络，是不清晰的。

解决方案之一是使用一个 HTTP 代理，为来自外部网络的请求提供透明的安全，同时不会过载 802.15.4 网络以及存在于其上的微型设备。在未来，可能有必要开发适合于这个域的更通用的安全机制，接下来构建一个真正的物联网，像因特网上的任何其他节点一样在因特网上是可见的。

8.12 物联网的应用

本节列出针对物联网的一个应用基础集合，可在 LoWPAN 和 IETF 协议上进行实现。

8.12.1 智能电网

一个智能电网使用双向数字通信技术，提供从提供商到消费者的平滑的和高效的电力交付，以便节省能源、降低成本和增加可靠性。世界各地的多个政府正在倡导智能电网，作为解决全球变暖和应急恢复问题的一种方式。

传感器和设备装备有 LoWPAN 能力，并被放置在各地，从发电站开始，经电力传输、电力分配并一路到消费者设备。这有助于电力设施实时地监测电力消耗，并以成本有效的方式管理它们的发电，方法是使用各种电力来源，包括绿色能源。

8.12.2 工业监测

LoWPAN 的工业监测应用可被关联到工程设施和制造厂中增加生产率、能源效率和工业操作安全的大范围方法。许多公司当前使用耗时和昂贵的人工监测法来预测故障，并调度维修或替换，以便避免成本高昂的制造停机时间。LoWPAN 可廉价地安装，提供更频繁的和更可靠的数据。LoWPAN 的部署可降低设备停机时间，并去除实施起来成本高昂的人工设备监测。另外，可将数据分析功能放入到网络中，去除对人工数据传递和分析的需要。工业监测可粗略地分成如下应用领域：过程监测和控制、机器监控、供应链管理和资产跟踪以及存储监测。

1. 过程监测和控制

这涉及将高级能源计量和细粒度计量技术与无线传感器联网组合使用，以便优化工厂操作、降低峰值需要，最终降低能源的成本，避免机器停机时间，并增加操

作安全。

一个工厂的监测边界经常不会涵盖整个设施，而是仅监测对过程而言，人们认为是关键的那些区域。容易安装的无线连接，将这条线扩展到包括周边区域和过程测量，以前采用有线连接是不可行的或实践中不能做到的。

2. 机器监控

机器监控是确保产品质量以及高效的和安全的设备操作。诸如振动、温度和电子签名等关键设备参数，针对异常进行分析，这些异常是即将发生的设备故障的暗示。

3. 供应链管理和资产跟踪

对于零售产业，在法律上负责所售商品的质量，就温度方面对不合适存储条件的早期检测，将降低从销售渠道中清除产生的风险和成本。例子包括集装箱发货、产品识别、货物监测、分配和物流。

4. 存储监测

传感器系统可被设计为防止受控物质扩散到地下水、地表水和土壤中。这个应用领域也可包括存储设施或其他基础设施（如管道）的盗窃/篡改防御系统。

8.12.3 结构监测

设施管理中的智能监测可使架构状态的安全检查和周期性监测是高效的。主供电节点可在构造的设计阶段就包括在内，或之后添加由电池装备的节点。所有节点是静态的和人工部署的。对于安全防护（如正常的室温），一些数据是不太紧急的，但事件驱动的紧急数据必须以一种非常紧急的方式加以处理。

8.12.4 保健

人们设想 LoWPAN 会被大量用于保健环境。通过去除烦人的导线而方便新服务的部署，并简化在医院和家庭看护中患者看护，它们具有巨大的潜力。在保健环境中，延迟的或丢失的信息可能是生死问题。

各种系统，范围从简单的用于远程辅助的可穿戴远端控制或带有监测各种指标的可穿戴传感器节点的中间系统，到研究生命动态的比较复杂的系统，可得到 LoWPAN 的支持。

8.12.5 连接的家庭

“连接的”家庭，或“智能”家庭，无疑是 LoWPAN 可被用来支持日渐增加数量之服务的一个领域：

1）家庭保安/安全。

2）家庭自动化和控制。

3）保健（见前一节）。

4）智能仪表和家庭娱乐系统。

在家庭环境中，典型情况下，LoWPAN 由数十个并可能在近期的未来由数百个不同特征的节点组成：传感器、执行器和连接的物体。

8.12.6 远程测量

在智能运输系统中，LoWPAN 扮演一个重要的角色。被集成在道路、车辆和交通信号中，它们对运输系统安全的改进有所贡献。通过交通或空中质量监测，就交通流优化方面，它们增加了各种可能性，并有助于降低道路堵塞。

8.12.7 农业监测

准确的时间和空间监测，可显著地增加农业生产率。由于自然限制，诸如一名农民不能在一天的所有时间都检查作物，或不充足的测量工具，在收成方面，运气经常扮演太大的成分。使用以战略性放置的传感器的一个网络，则在没有劳动密集型现场测量的情况下，可自动地监测各指示器（如温度、湿度）和土壤条件。例如，传感器网络可实时提供有关作物的精确信息，使事务可降低水量、能量和杀虫剂的使用并增强环境保护。传感数据可被用来为农场找到最优环境。另外，有关农场条件的数据可由传感器标签存储，这可用在供应链管理中。

参 考 文 献

1. “IEEE Std. 802.15.4-2006—IEEE computer society, part 15.4: wireless medium access control (MAC) and physical layer (PHY) specifications for low-rate wireless personal area networks (WPANs),” September 2006.

2. “Transmission of Ipv6 packets over IEEE 802.15.4 networks,” RFC 4944, September 2007.

3. “Internet protocol, version 6 (Ipv6) specification,” RFC 2460, December 1998.

4. “Transmission of Ipv6 packets over Ethernet networks,” RFC 2464, December 1998.

5. “IP version 6 addressing architecture,” RFC 4291, February 2006.

6. “Neighbor discovery for IP version 6 (Ipv6),” RFC 4861, September 2007.

7. “IPv6 stateless address autoconfiguration,” RFC 4862, September 2007.

8. "Internet protocol, STD 5," RFC 791, September 1981.
9. "User datagram protocol, STD 6," RFC 768, August 1980.
10. K. Roemer and F. Mattern, "The design space of wireless sensor networks," December 2004.
11. "IPv6 over low power WPAN (6lowpan)," IETF's Internet Area Working Group.
12. "Constrained RESTful environments (core)," IETF's Applications Area Working Group.
13. "IPv6 over low-power wireless personal area networks (6LoWPANs): overview, assumptions, problem statement, and goals," RFC 4919.
14. "Transmission of IPv6 packets over IEEE 802.15.4 networks," RFC 4944.
15. "Compression format for IPv6 datagrams in 6LoWPAN networks (draft-ietf-6lowpan-hc-13)," Work in progress.
16. "Harbor research's pervasive Internet/M2M forecast report," 2009.
17. "Dynamic HOST CONFIGURATION PROTOCOL for IPv6," RFC 3315.
18. "Service location protocol, version 2," RFC 2608.
19. "An architecture for describing SNMP management frameworks," RFC 2571.
20. "Unicast-prefix-based IPv6 multicast addresses," RFC 3306.
21. "Embedding the rendezvous point (RP) address in an IPv6 multicast address," RFC 3956.

第 9 章　6LoWPAN：采用 IPv6 互联物体

Gilberto G. de Almeida, Joel J. P. C. Rodrigues, Luís M. L. Oliveira

9.1 引言

无线传感器网络（WSN）概念是在 20 世纪 90 年代开发的，并形成由数千个自治传感器节点组成的网状网[1]。这些空间分布的传感节点支持各种军事、民用和工业应用，包括环境条件的实时监测、安全、监控、资产跟踪和建筑自动化。

在设备非常昂贵并使用专用协议进行通信的时代，完成互操作和应用开发是非常困难的。流行的协议，像在 1998 年引入的 ZigBee，通过支持不同设备间的通信，而帮助将这个领域中的研究推进了一步。

在 2003 年 5 月，针对低速率无线个域网的 IEEE 802.15.4 无线媒介访问控制（MAC）和物理层（PHY）规范标准[2]的引入，建议了一个通用的物理和媒介访问控制（MAC）平台，它支持来自不同制造商的硬件间的通信。传感器硬件的价格下降，与 2007 年 9 月“RFC 4944——在 IEEE 802.15.4 网络上传输 IPv6 分组”[3]规范的发行相结合，为 WSN 概念的实现创造条件。

低功率无线个域网（LoWPAN）是由小型、自治、可编程设备组成的网络。WSN 是一种类型的 LoWPAN，目标主要在于物理环境参数的监测和执行。每台设备，典型情况下是电池供电的，都装备有一个无线电数据链路，并与传感器和执行器有接口。LoWPAN 通常基于通信栈的物理和数据链路层的 IEEE 802.15.4 标准。

LoWPAN 节点的特征为低功率、低数据速率（在 20kbit/s 和 250kbit/s 之间），并具有有限的处理和存储容量。结果，它们在大量时间段都必须睡眠以便节省能量，通常适合短距离低功率无线电。由于干扰和其他环境因素，设备之间的通信是丢失性的。

需要一种新的范型，支持低功率设备参与到因特网。“物联网”范型[4]出现了，其中所有嵌入式设备和网络都以原生方式支持 IP，连接因特网，这独立于所用的物理和 MAC 层协议。

“物联网”被看作是因特网的最大挑战和机遇[4,5]。2008 年，许多业界领导者，倡导使用“物联网”，形成 IP 智能物体联盟[6,7]，因特网工程任务组（IETF）设立了“LoWPAN 上 IPv6”章程[8]。

人们认为 IPv6 比 IPv4 更适合 LoWPAN[9]，原因是它提供大得多的寻址空间、比较容易的应用开发过程和较佳的自动配置机制。但是，IPv6 协议的设计，没有

考虑工作在 LoWPAN 设备的约束之下的情况。

引入 IPv6LoWPAN（6LoWPAN）规范，在 IEEE 802.15.4 LoWPAN 之上支持使用 IPv6 协议，方法是在它们之间创建一个绑定层。这个层，即 6LoWPAN 适配层，处理分组分片和压缩。

WSN 适合于部署在大量的、各种物理环境上，形成自组织无结构的网状网络。6LoWPAN 节点的这种分布式特征使网络更具扩展能力和更加鲁棒，但也要求一种网络自动配置的方法。

邻居发现（ND）协议[10]是将这种能力提供给常规 IPv6 网络的机制，处理地址指派以及邻居和路由器发现。但是，考虑到 LoWPAN 节点的物理约束，例如在层 2 缺乏组播支持、长的睡眠时段和需要保留能量，6LoWPAN 上的 ND 要求一种不同的方法，焦点是可用能量的高效使用。

人们正在建议"6LoWPAN 的 ND 优化"[11]，迎合 WSN 的非常特定的需要，目标是降低能量使用并延长网络寿命。

本章将讨论低功率无线网络和设备、IEEE 802.15.4 标准以及 6LoWPAN 规范和适配层，包括建议的 6LoWPAN ND 优化。最后，它将提供在 TinyOS 和 ContikiOS 开源操作系统上的 6LoWPAN ND 支持概述。

9.2 传感器节点

LoWPAN 由自治低功率传感器节点、小型计算设备（装备有复合 IEEE 802.15.4 标准的无线电接收转发器）组成。它们通常是小型的、低成本的、低功率设备，它们可包括许多传感器（也称作变换器）或执行器。

这些特征使 LoWPAN 节点成为在广域上以大量方式（作为一个网状 WSN 的组成部分）监测和部署的一个良好选择。

每个传感器节点通常是由一块低容量的电池供电的，并有一个计算设备［正常情况下是一个微控制器或一个低功率中心处理单元（CPU）］、少量内存和某种类型的永久存储（通常是闪存）。一个板上无线电接收转发器支持节点之间的通信和与汇聚节点（网关或边缘路由器）的通信。设备可装备有光、温度、湿度、声的传感器，以及像光发射二极管（LED）、电动机或扬声器等执行器。图 9.1 给出了一个无线传感器节点的最基本组件。

传感器节点是任何 WSN 的基本组件，它们提供如下基本功能[1,12,13]：

1）用于不同传感器的信号调整和数据获取。

2）所获取数据的临时存储。

3）数据处理。

4）用于诊断的被处理数据的分析，以及可能的告警生成。

5）自监测（如供电电压）。

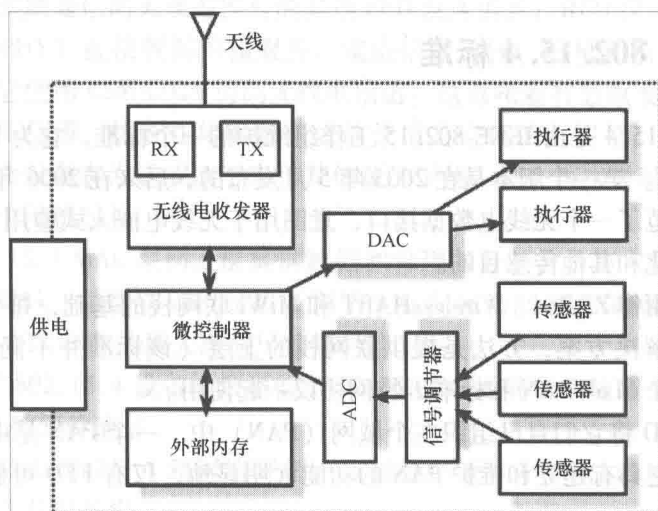


图9.1　传感器节点硬件架构

6）测量任务的调度和执行。

7）传感器节点配置的管理。

8）数据分组的接收、传输和转发。

9）通信和联网的协同及管理。

为提供上述功能，如图9.1所示，一个传感器节点由一个或多个传感器、一个信号调整单元、一个模拟-数字转换（ADC）模块、一个CPU、内存、一个无线电收发器和一个能量供电单元组成。取决于部署环境，以一个合适的封装保护传感器硬件不受机械和化学侵入是必要的。

传感器节点也称作“微尘”，这是由美国加利福尼亚大学的研究人员创造的一个术语。各节点通常被分类为精简功能设备（RFD）或全功能设备（FFD）。前者是自治的和电池供电的，具有非常有限的计算资源。另外，后者通常用作多跳网络中的网关或路由器。

短距离无线电意味着一些节点必须作为处于无线电范围之外邻居节点之间的转发器。对于网络的完整性而言，路由节点是重要的，所以必须采用能量节省战略，以便延长节点和网络本身的寿命。

对于LoWPAN，自组织是非常重要的。为形成一个LoWPAN，各节点必须能够自配置、定位邻居节点、建立路径和发现网关路由器。LoWPAN的另一个重要方面是冗余和容错。网络必须能够处理并路由绕过失效的节点（通过多条路径）。

同样，考虑到一个LoWPAN周围的变化的环境，网络持续地确定最佳路径以便优化其性能就是必要的。

9.3 IEEE 802.15.4 标准

IEEE 802.15.4 是由 IEEE 802.15 工作组维护的一个标准，它为 LoWPAN 规范物理和 MAC 层。第一个版本是在 2003 年 5 月发布的，后来在 2006 年和 2007 年得到更新。它规范了一个无线电数据接口，意图用于无线电嵌入式应用，如建筑自动化、工业自动化和其他传感目的。

该标准也用作 ZigBee、WirelessHART 和 MiWi 联网栈的基础，每个栈都提供一个完整的联网解决方案，方法是提供联网栈的上层（该标准并不涵盖这层）。另外，它可与一个 6LoWPAN 和标准因特网协议一起使用。

FFD 和 RFD 将它们自己组织在个域网（PAN）中。一个 PAN 是由一个 PAN 协调器控制的，它具有建立和维护 PAN 的功能（明显地，仅有 FFD 可假定具有 PAN 协调器的角色）。

IEEE 802.15.4 MAC 提供两种操作模式，即异步无信标模式和同步信标支持模式。无信标模式要求节点在所有时间都侦听其他节点的传输，这可能快速地耗光电池功率。支持信标的模式被设计为支持发送器和接收器之间的信标分组的传输，提供节点间的同步。在支持信标的模式中，PAN 协调器广播包含有关 PAN 之信息的一个周期性信标。由信标提供的同步，支持设备在传输之间睡眠，这得到能源效率和扩展的网络寿命。

在支持信标的模式中，两个连续信标之间的时段定义了可分成 16 个槽的一个超帧结构。一个信标总是占据第一个槽，而其他槽则用于数据通信。在这些槽中，采用冲突避免的槽式载波侦听多路访问（CSMA/CA）被用于数据通信。为支持低延迟应用，PAN 协调器可预留一个或多个槽，由有保障的时槽指定，时槽被指派到运行这种应用的设备。在这种情形中，这些设备不需要使用基于冲突的媒介访问机制。在无信标的模式中，没有超帧结构，且不能预留确保的时槽。结果是，仅有随机访问方法，如无时槽的 CSMA/CA，可被用于媒介访问。

一个 PAN 可采用如下两种网络拓扑之一[1]：

1）星形拓扑（见图 9.2）：使用一个主从网络模型，其中一个 FFD 假定具有 PAN 协调器的角色，并控制所有的网络操作；其他节点可以是 RFD 或 FFD，并进而与 PAN 协调器通信（这个拓扑比较适合小型网络）。

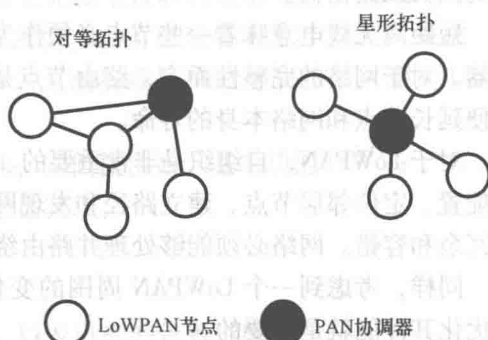


图 9.2 LoWPAN 拓扑图示（对等和星形）

2）对等拓扑（见图 9.2）：一个 FFD 可与无线范围内的其他 FFD 通

信，并可使用多跳通信向无线范围外的其他 FFD 发送消息；RFD 仅可与 FFD 通信。

物理层（PHY）提供数据传输服务，实施信道选择、能量和信令管理并控制分组数据流。它使用 CSMA/CA 访问无线电信道。这意味着有数据要发送的一个无线电将首先侦听信道，检查在传输数据之前信道是空闲的。但是，如果信道是忙的，由于另一个设备在发送或由于来自其他源的干扰，则在再次尝试之前，无线电将发送延迟一个随机时长。

IEEE 802.15.4 MAC 层向上层提供数据和管理服务。它负责 PAN 关联和去关联、帧验证和确认、信道访问机制以及用于支持信标访问的有保障时槽和信标管理。MAC 帧将 127 字节的最大净荷尺寸提供给上层。

所有 IEEE 802.15.4 设备携带类似于 Wi-Fi 或以太网卡中使用的 MAC 地址的一个唯一 64 比特硬件地址。

但是，为减少首部尺寸和网络开销，设备被指派一个 16 比特的短本地地址，这实际上缩短了分组长度。

在最左比特为零的短 16 比特地址，表明单播地址，为实际地址留下 15 比特。以100（二进制）开始的地址是组播地址，这留下一个 13 比特的地址空间。地址 0×FFFF 和 0×FFFE 用于广播。

802.15.4 标准是非常灵活的，并支持多种拓扑，包括星形和对等拓扑。拓扑选择是应用相关的。例如，星形拓扑提供低延迟，但对等网络提供较宽的覆盖范围。图 9.2 形象地给出了对等和星形拓扑。

LoWPAN 帧

在 LoWPAN 帧中封装的数据报带有 LoWPAN 首部栈的前缀。每个首部由一个首部类型字段（1 字节），后跟零个或多个首部字段，这取决于类型。

第一个首部总是 LoWPAN 指派。这个首部（1 字节长）指明该帧是一个 LoWPAN帧（或不是），并声明后续首部的类型。

该标准定义了处理网状网、广播和分片的 LoWPAN 特定首部。当使用一个以上的 LoWPAN 首部时，正确的顺序是网状网寻址首部、广播首部和分片首部。

表 9.1 列出了指派字节的当前指派的比特模式。

表 9.1　LoWPAN 首部类型

比特模式	首部类型
00 ××××××	NALP——不是一个 LoWPAN 帧
01 000001	IPv6——非压缩 IPv6 首部
01 000010	LOWPAN_HC1——压缩的 IPv6 首部
01 010000	LOWPAN_BC0——广播首部
01 111111	ESC——后跟额外的分派字节

（续）

比特模式	首部类型
10 ××××××	MESH——网状网首部
11 000×××	FRAG1——第一个分片首部
11 100×××	FRAGN——接下来的分片首部

9.4 6LoWPAN

在IEEE 802.15.4低功率无线链路之上传输IPv6分组的方法，被称作6LoWPAN，这是低功率无线个域网之上传输IPv6的一个缩略语。这是IETF工作组基于所陈述问题（RFC 4919）[14]设计规范的名称（RFC 4944）[3]。

6LoWPAN规范描述在IEEE 802.15.4之上传输IPv6分组的帧格式、链路层地址形成、首部压缩和无状态地址自动配置。

考虑LoWPAN节点的受约束能力，IPv6协议支持实现起来是有挑战的。但是，支持IPv6，则支持针对网络部署、配置、管理和调试而使用尝试和测试工具；提供大量地址空间；支持到其他IPv6网络（包括因特网）的无缝连接；用于比较简单的应用开发。

9.4.1 6LoWPAN适配层

6LoWPAN规范，在链路和物理层依赖IEEE 802.15.4标准，在网络层依赖IPv6。但是，IEEE 802.15.4标准没有完全满足IPv6协议的需求。

针对一条IPv6分组，最小支持的最大传输单元（MTU）是1280字节。但是，由IEEE 802.15.4提供的帧尺寸仅有127字节长，其中在考虑链路层额外负担之后，仅有81~102字节可用于净荷。

一条完整的IPv6分组通常不能封装在一个IEEE 802.15.4帧中。一个简单的IPv6首部占40字节，仅为额外的首部和上层用途留下41字节。在为一个用户数据报协议（UDP）首部预留8字节或为一个传输控制协议（TCP）首部预留20字节之后，仅为实际的应用数据留下不多的字节可用。

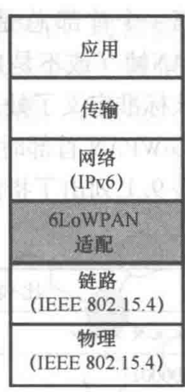


图9.3　6LoWPAN适配层

因此，为调整网络层对链路层的需求，6LoWPAN工作组提出创建一个适配层。这个适配层，如图9.3所示，位于层2和层3之间，实际上将网络层与链路层解耦开来。

6LoWPAN适配层提供几项关键功能，优化IPv6分组到IEEE 802.15.4帧的

映射：

1）分组分片：IPv6 分组（不能封装到单条 LoWPAN 帧中）的分片和重组。

2）首部压缩：冗余首部数据压缩，其中使用一种无状态方法，降低网络开销并增加数据吞吐量。

3）无状态地址自动配置：通过绕过 IEEE 802.15.4 组播能力缺乏的问题，支持 IPv6 地址自动配置。

4）ND：能量高效邻居和路由器发现。

三种不同的 LoWPAN 架构类型定义如下：①自组织 LoWPAN，没有基础设施；②简单的 LoWPAN，带有一个边缘路由器；③扩展的 LoWPAN，带有多个边缘路由器。这三个不同的 6LoWPAN 架构类型如图 9.4 所示。

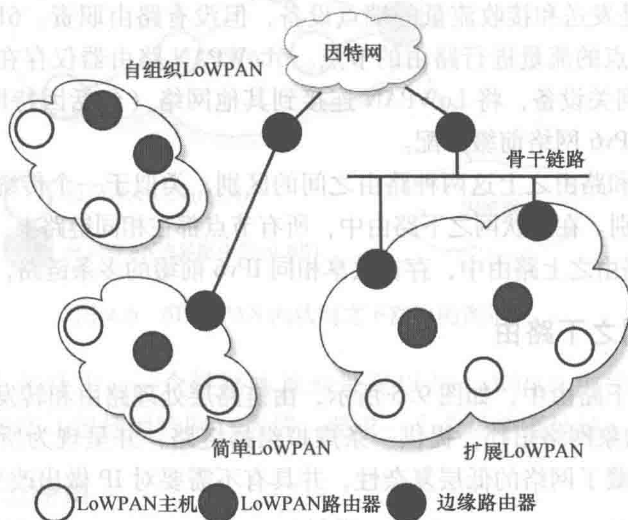


图 9.4 LoWPAN 类型图示

9.4.2 6LoWPAN 路由

人们为 LoWPAN 路由提出了多个协议[8,15]。这些路由机制考虑到网络目的和架构需求。LoWPAN 中的研究工作导致几个能量感知路由协议的开发，其中网络寿命最大化是主要关注的问题[15]。从网络结构观点，将 LoWPAN 路由协议分为三类：扁平路由、层次路由和位置路由。在扁平路由中，所有节点具有相同角色或功能。在层次路由中，不同节点在网络上扮演不同角色：具有较多资源（能量、计算能力和内存）的节点可被用在多跳转发中，而其他节点可用于感知操作。在位置路由中，传感器节点以它们的位置进行寻址，而路由数据时使用节点位置。节点位置可采用所接收信号的强度或使用全球定位系统（GPS）加以估计。

LoWPAN 基于多跳转发，原因是个体节点经常缺乏到达目的地的无线电范围。

中间节点必须在源和目的地之间转发分组。

转发是在输入接口上接收一条分组并在输出接口上发送该分组的过程。因为多数设备具有单个接口，则它被用作这两个目的。转发经常由协议栈的低层进行处理。

路由是这样一个过程，它使用一种路由协议确定一条分组的最佳路径。

在一个 LoWPAN 中路由和转发可以三种不同的方式完成：链路层网状网之下（mesh-under）、6LoWPAN 网状网之下和路由之上（route-over）[9]。

6LoWPAN 上的路由可在链路层（网状网之下）或在网络层（路由之上）实施。如将看到的，这两种方法都有优势和缺陷。

6LoWPAN 上的设备可被分类为节点（6LN）、路由器（6LR）或边界路由器（6LBR）。6LN 是发送和接收流量的端点设备，但没有路由职责。6LR 是另外的将目的地为其他节点的流量进行路由的节点。6LoWPAN 路由器仅存在于路由之上拓扑中。6LBR 是网关设备，将 LoWPAN 连接到其他网络（包括因特网），也处理在 6LoWPAN 上的 IPv6 网络前缀分配。

网状网之下和路由之上这两种路由之间的区别，类似于一个传统以太网上网桥和路由之间的区别。在网状网之下路由中，所有节点都在相同链路上，由一个或多个 6LBR 服务。在路由之上路由中，存在共享相同 IPv6 前缀的多条链路，由 6LR 互联。

9.4.3 网状网之下路由

在网状网之下路由中，如图 9.5 所示，由链路层处理路由和转发。网状网之下路由从 IPv6 层抽象网络拓扑，提供一条虚拟组播链路，并呈现为所有节点都是直接可达的。这隐藏了网络的低层复杂性，并具有不需要对 IP 做出改变的优势。

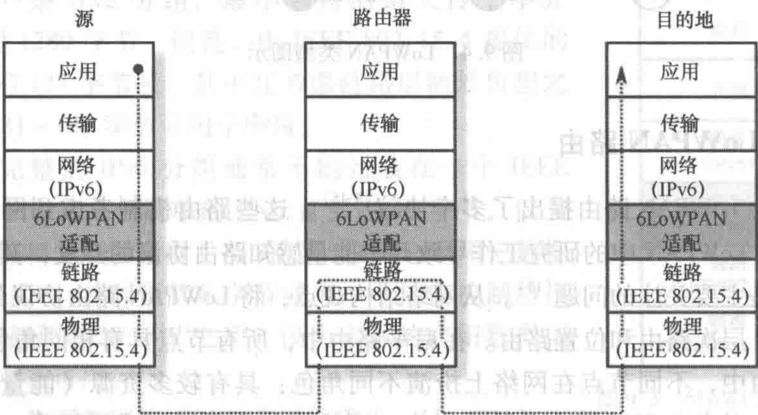


图 9.5 网状网之下路由的图示

在网状网之下模式中，分组分片可在多跳上交付，使用多条路径，为在网络层做出路由判定，这排除了在每个无线电跳重组分组的需要。

网状网之下模式的最大问题是，不可能使用标准的网络诊断工具，像 ping 或 traceroute，来诊断故障并分析性能，原因是所有节点似乎都是一跳远的[16]。

图 9.6 给出了一个网状网之下的 6LoWPAN。所有节点都能够代表另一个节点来转发分组，对 IPv6 层看来就像它们都只有一跳远。边界路由器将 6LoWPAN 连接到其他网络。

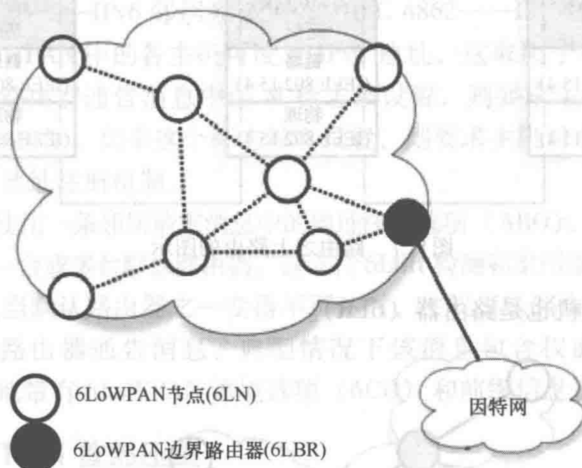


图 9.6　6LoWPAN 网状网之下路由的图示

在网状网之下拓扑中，一个链路本地地址足以与 LoWPAN 内的各节点通信。但是，为与其他网络（包括因特网）通信，要求一个全局单播地址。

9.4.4　路由之上路由

在路由之上模式中，路由决策是在网络层实施的，而分组转发发生在较低层。图 9.7 给出了路由过程。

出于简单性，每个路由之上网络都共享单个全局 IPv6 前缀，6LR 转发分组到一台默认路由器或使用一种逐跳路由协议，如低功率和丢失性（RPL）网络的路由协议[17]。

路由之上路由方法的一个重要缺陷是，与网状网之下路由方法不同，每个无线电跳也是一个 IP 跳，所以在每个节点，分组分片和重组都是必要的。同样，因为层 3 处理路由过程，这意味着分片不能在多个路径上交付。

在 IP 层处的路由协议必须也能够查询硬件参数（无线电范围、可用能量等），为一条给定分组建立最佳路由。

在上面，路由之上路由方法支持使用高级路由算法，并针对性能评估和网络排错而使用调试和测试网络工具。

图 9.8 形象地给出了一个典型的路由之上 LoWPAN。注意，与网状网之下路由

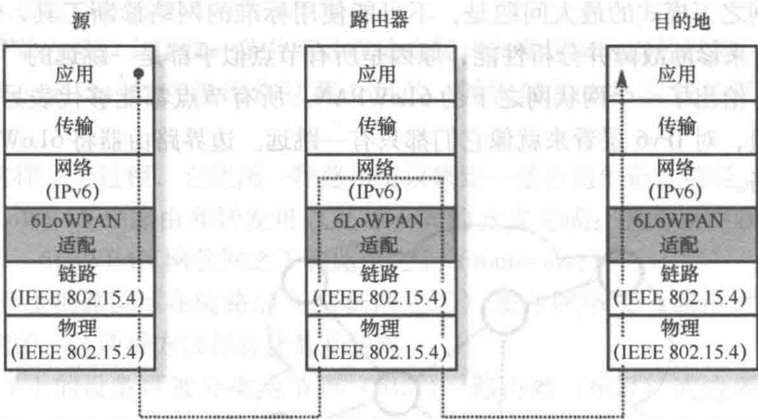


图 9.7　路由之上路由的图示

方法不同，一些主机也是路由器（6LR）。

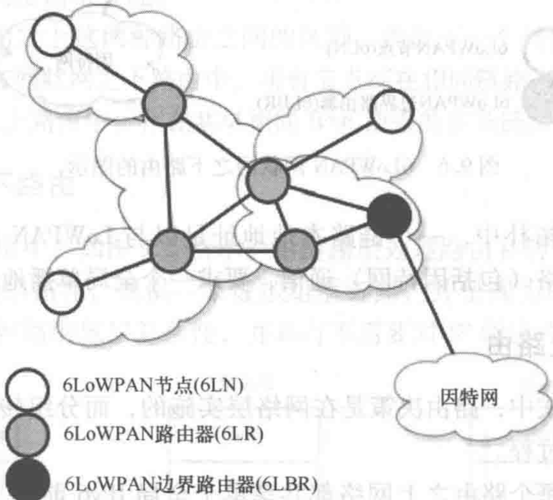


图 9.8　6LoWPAN 路由之上路由的图示

当使用路由之上路由方法时，本地链路地址支持与直接无线电链路节点的通信，但为了与多跳远的设备通信，要求使用全局地址。

9.4.5　6LoWPAN 地址指派

IEEE 802.15.4 标准定义两种设备寻址模式，它支持使用全尺寸 IEEE 64 比特扩展地址（EUI-64）或使用另外一种 16 比特短地址，在 PAN 内是唯一的，在一个设备关联到 LoWPAN 之后指派的。

为加入一个 6LoWPAN，各设备必须有一个有效的 IPv6 地址。IP 地址可以是人工

指派的或自我指派的。后者是最实用的、可扩展的方法，原因是它不要求人类干预。

地址自动配置协议可以是有状态的或无状态的。动态主机控制协议（DHCP）是有状态协议的一个例子。DHCP 服务器保持一个地址租期的列表，这代表网络的状态。

另外，ND 协议提供无状态地址自动配置。基于由路由器通告的网络前缀，每台主机产生自己的 IPv6 地址。地址的唯一性可使用重复地址检测（DAD）加以验证。

有状态和无状态地址指派协议可在一个 6LoWPAN 上共存和相互补充。

依据 RFC 4861——IPv6 邻居发现[18] 和 RFC 4862——IPv6 无状态地址自动配置[19]，一个 6LoWPAN 中的各主机可配置 IPv6 地址，这取决于所接收到的路由器通告消息。如果路由器通告消息中的 M 标志被设置，则要求主机使用 DHCPv6 进行非 EUI-64 地址指派。如果这个标志没有设置，则要求主机对非 EUI-64 地址实施 DAD，其中使用地址注册机制。

要求各主机使用一条邻居请求消息中的地址注册选项（ARO），将非链路本地 IPv6 地址注册到它的一台或多台默认路由器。这支持 6LBR 检测和重用重复地址请求。

在启动时或当默认路由器之一变得不可达时，主机发送路由器请求消息。各主机从 6LBR 接收路由器通告消息，典型情况下该消息包含权威边界路由器选项（ABRO），可选地带有 6LoWPAN 语境选项（6CO）和前缀信息选项（PIO）。

9.4.6　6LoWPAN 首部压缩

IPv4 协议或 ZigBee 栈定义单个首部，与此做法不同，一个 6LoWPAN 像原始 IPv6 协议一样，使用堆叠式首部。当一台设备直接向另一个节点发送消息时，对网状网联网，它不使用非必要的首部字段或分片，并仅使用最少的必要首部。在最简单的情形中仅使用分片和压缩首部。在每个首部的开始部分，一个首部类型字段识别首部格式。

在 RFC 4944 中定义了 6LoWPAN 首部压缩。它定义了一种无状态压缩方案，由两部分组成，即首部压缩 1（LOWPAN_HC1）和首部压缩 2（LOWPAN_HC2）。HC1 目标是压缩 IPv6 首部，而 HC2 支持 UDP 首部的压缩。

IEEE 802.15.4 首部仅包含下一跳的源和目的地址。如果一条分组应该被发送到这样一个节点，它不是源的一个邻居，则需要另外一个协议来实现这项功能，如 IEEE 802.15.5。使用 IPv6，源和最终接收者地址都被包括在 IPv6 逐跳选项首部之中。此外，使用压缩首部，这个信息可能丢失。这个问题的解决方案是引入网状网首部，它被用来支持层 2 转发。

6LoWPAN 规范为在 6LoWPAN 中交付 IPv6 分组，定义了一种无状态的 IPv6 首部压缩格式。

在没有显式存储任何压缩语境状态的条件下，压缩 IPv6 首部是可能的，原因是各主机共享相同的 6LoWPAN。LOWPAN_HC1 编码是链路本地通信的一种经优化的 IPv6 首部压缩方案，它依赖于有关链路的信息，取得压缩效果。多数 IPv6 首部

字段，如IPv6长度字段和IPv6地址，可从一个分组中清除，做法是假定适配层可从链路层帧中的首部重建它们。

适配、网络和传输层中的首部字段通常携带常见值。为降低传输开销，首部压缩被用来将那些首部字段压缩到数比特，同时当必要时，保留表示非压缩字段的一种方式。

人们期望如下IPv6首部值在6LoWPAN上是常见的，所以构造了HC1首部以便高效地压缩它们：

1）版本字段总是IPv6。

2）IPv6源和目的地址都是链路本地的。

3）源或目的地址的接口标识符可从层2源和目的地址中推断得到（如果是从一个下层802.15.4 MAC地址推断的话）。

4）分组长度可从层2推断得到。

5）流量类和流标签字段总是为零。

6）下一首部字段是UDP、因特网控制消息协议（ICMP）或TCP。

在IPv6首部中总是需要完整携带的唯一字段是跳限制字段（8比特）。

与这种常见情形不同的值将必须在线路中携带。但是，如果首部匹配这种情形，则可从40字节压缩到2字节（1字节用于HC1编码，1字节用于跳限制字段）。

9.4.7　6LoWPAN分片

一个6LoWPAN使用分片首部支持IPv6对低层MTU所要求的最小值（这是1280字节）。无论何时净荷太大，不足以封装到单个IEEE 802.15.4帧时，它就被分片成几条分组。

分组被分成链路层分片并被发送。第一个分片必须前面带有第一分片首部。后续分片包括一个数据报偏移字段，它支持分片的重新组装。图9.9形象地给出了第一个和后续分片首部的格式。

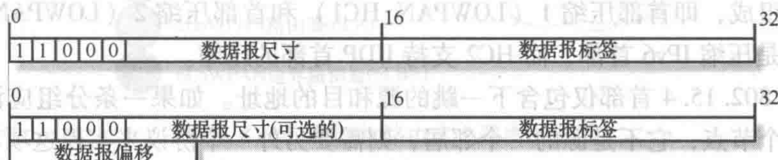


图9.9　第一个和后续分片首部

数据报尺寸字段必须采用第一个分片发送，并可在后续分片上忽略。但是，可将其包括在内，以防止在乱序接收情形中方便目的地的分片重新组装。除了最后一个分片的尺寸外，其他分片的尺寸必须是8的整数倍。

9.4.8　6LoWPAN邻居发现

针对低功率和丢失性网络的ND优化（draft-ietf-6lowpan-nd-17）[11]是IETF

6LoWPAN 工作组的一项进行中的建议规范。当前处在修订版 17，其目标是更新 RFC 4944[3]，如果得到批准的话。它描述针对低功率网络对 IPv6 ND、寻址机制和 DAD 的简单优化。

虽然根据预期，标准 IPv6 ND 协议应该工作在 6LoWPAN 上，但实现 6LoWPAN ND 优化存在紧迫的原因。IPv6 ND 协议不是针对非中转型无线链路设计的，它大量使用组播，这在丢失性、低功率网络中是低效的和不现实的。IPv6 ND 假定本地链路节点总是单跳远的，组播是可用的，各节点总是在侦听，但在 LoWPAN 中情形却不是这样的。

虽然 IEEE 802.15.4 标准在链路层支持广播，但由于能量保护策略，它有限地使用组播信令。设计 6LoWPAN ND 优化是为了解决这些问题。通过以地址注册替换地址解析，它们简化了 ND 信令。通过提供主机发起的对路由器通告的请求，它们也去除了对周期性路由器通告的需要。同样，在多数情形中，它们能够将组播消息优化为单播消息。

6LBR 在 6LoWPAN 中扮演了一个重要角色。它负责在 LoWPAN 间传播 IPv6 前缀和首部压缩语境信息。6LBR 也维护主机 IPv6 地址和 EUI-64 标识符的一个网络范围缓存，这使它能够检测和避免重复地址。另外地，DHCPv6 可被用来确保地址在网络上是唯一的。

6LoWPAN ND 假定每个 IPv6 是从唯一 EUI-64 地址派生得到的，所以缺省情况下，它不要求 DAD 或地址解析。但是，该规范不支持这种地址的一个可选多跳 DAD 机制，这些地址不是从 EUI-64 地址派生得到的。

这些优化导致局部网络中信令消息的显著降低，得到显著的能量节省，扩展了网络的寿命。

为取得这些目标，该规范定义三个新的 ICMPv6 消息选项：必备的 ARO、可选 ABRO 和 6CO。

也定义了两种新的 ICMPv6 消息类型，实施可选的多跳 DAD，即重复地址请求（DAR）和重复地址确认（DAC）。

图 9.10 展示出一条周期性的组播路由器通告是如何为主机发起的交互通信替换的。主机发送一条路由器请求（RS）消息［包括源链路层地址（SLLA）］，从而路由器可以一条单播路由器通告（RA）消息做出应答。RA 消息可包括 SLLA、ABRO、6CO 和 IPv6 PIO。

地址注册过程如图 9.11 所示。主机发送一条单播邻居请求（NS）消息到路由器，带有 ARO。路由器以带有 ARO 和注册状态的一条单播邻居通告（NA）消息做出应答。状态指明一次成功的注册，或一次失效（由于重复的地址或因为路由器的注册缓存满了）。

可选的多跳 DAD 过程如图 9.12 所示。它可被用于路由之上的网络，在 6LoWPAN 内为基于非 EUI-64 的地址确保地址唯一性。这类似于标准地址注册过程，例外情况是，6LBR 负责管理地址注册缓存，主机尝试与其注册的中间路由

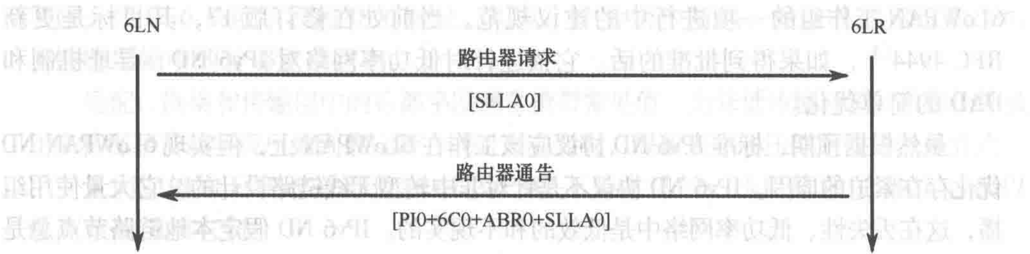


图 9.10 主机发起的路由器发现

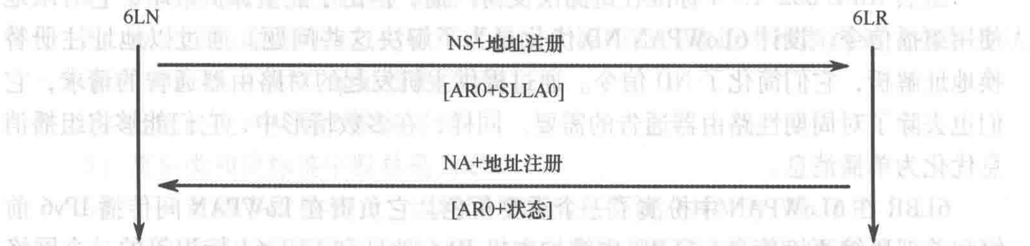


图 9.11 主机地址注册

器，必须首先在 6LBR 中检查地址是否是一个重复地址。这是使用新 DAR 和 DAC ICMPv6 消息做到这一点的。

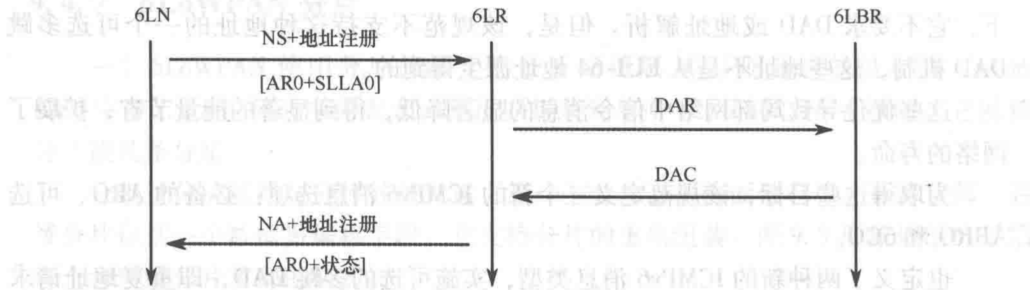


图 9.12 带有多跳 DAD 的主机地址注册

9.5 6LoWPAN 实现

存在可用于部署一个 6LoWPAN 的几个操作系统。用于 WSN 的两个最流行的开源嵌入式操作系统是 TinyOS[20] 和 ContikiOS[21]。

9.5.1 TinyOS

TinyOS[22] 是特别为 WSN 设计的一个嵌入式的和开源的操作系统。它是一个模块化的、单栈、协作式多任务系统，是以网络嵌入式系统 C（nesC）编程语言编写的。它支持非阻塞输入/输出（I/O），基于由编译器引入的异步回调做到这一点。

这个操作系统是自 2000 年以来开发的，是加利福尼亚大学伯克利分校、英特

尔研究院和 Crossbow 技术工作室协作研究的成果，已经成长为一个国际性的社团，即 TinyOS 联盟。它主要由学术研究人员和硬件厂商使用、开发和支持。

它是在厂商友好的伯克利软件分发（BSD）许可证下发放许可的，这使任何人可拷贝、修改、分发或销售源代码。

TinyOS 的网络栈被称作伯克利低功率 IP（BLIP）栈，并实现许多基于 IP 的协议。BLIP 当前不是完全兼容标准的。但是，它确实提供相当的互操作能力。一个重新编码的、更符合标准的栈，称作 BLIP-2.0，当前正在开发，但还没有达到稳定状态。

BLIP 的 6LoWPAN 实现支持：

1）分组分片和重组。

2）IP 和 UDP 首部的压缩和解压缩。

3）对 ICMP 回声请求的响应。

4）处理 UDP 之上的通信。

5）剖析网状网和广播首部。

但是，它有几项缺点和缺失的特征：

1）它使用所谓的“主动消息”，封装 6LoWPAN 净荷。这意味着，802.15.4 净荷带有一个 1 字节 AM 类型字段作为前缀，这防止与其他操作系统的互操作。

2）设备的 EUI-64 不被用来产生 IPv6 地址。

3）不支持网状联网和多跳。

4）没有实现 ND，所以要使用链路本地广播。

5）UDP 端口号压缩还没有得到实现。

9.5.2 ContikiOS

ContikiOS[15] 是用于内存高效的、联网的、嵌入式系统和 WSN 的一个高移植的、多任务型操作系统。它源于 2003 年瑞士计算机科学研究所，目前由来自工业界（包括 SAP、思科和 Atmel）和学术界的一群开发人员进行开发。它已经被移植到 20 多个平台。

ContikiOS 是一个抢占式的、多线程、实时操作系统。像 TinyOS 一样，在 BSD 许可证下可得到 ContikiOS，由此是免费的和开源的。与 TinyOS 不同的是，它是使用一个比较简单的工具链单纯由 C 编程的。它使用协议线程和动态代码载入来支持多线程，这在不必重新载入操作系统的条件下，载入各项应用。

存在两个网络栈，一个是 Rime 低功率无线电联网栈（用于 WSN 内的通信），另一个是 uIP（世界上最小的、经认证的嵌入式 IPv6 栈）。

Rime 栈实现针对 WSN 做过优化的各种网络协议，目标锁定在可靠的数据收集、尽力而为网络洪泛、多跳块式数据传递和数据传播。

uIP 联网栈支持 IPv4 和 IPv6，并包含一个 6LoWPAN 实现。该 IPv6 栈标记带有“IPv6 就绪”图标，这证明它是符合 IPv6 标准的。6LoWPAN 实现也在努

力做到符合标准，但还没有完成。IP分组是通过Rime栈在多跳路由上以隧道方式传输的。

ContikiOS可用在独立的微尘节点和基站微尘节点。基站微尘节点被连接到一台计算机，并就操作系统而言，该节点作为一个标准层2网络接口。

系统是极端模块化的，在不需要以闪存方式写入整个系统的条件下，支持各模块进行升级。空中编程，指一个网络中的所有节点可在仅数分钟内进行升级。

ContikiOS提供称为Coffee的一个基于闪存的文件系统，用于在每个微尘节点内存储数据，支持多个文件共存在相同物理板的闪存上。

ContikiOS的一项强大特征是开发和仿真环境。为方便软件开发和调试，ContikiOS提供基于Java的Cooja仿真环境。在不需要部署实际硬件的条件下，Cooja支持WSN的实际软件仿真。

ContikiOS的6LoWPAN实现支持：

1）分组分片和重组。

2）IP和UDP首部的压缩和解压缩。

3）802.15.4 16比特和64比特地址。

4）HC01压缩（draft-hui-6lowpan-hc）。

在6LoWPAN层没有实现ND；遵循RFC 4861，它是由IPv6层处理的。

9.6 小结

WSN作为一项技术，将逐步成熟，它将改变人们度量、理解和管理物理过程的方式。不同类型的数据和在不同位置收集的数据首次可合并在一起，并从任何地方进行访问。

可预测在近期几个日常物体将有一条因特网连接，这是“物联网”愿景。在所有小型物体中支持一个IP族，方便了同时进行的应用开发和连接到因特网。在智慧城市中，环境数据将通过因特网向市民提供有用的信息。例如，空气质量、运输信息、紧急服务等。但是，“物联网”部署远远落后于人们的期望，主要是因为难以部署新的应用，连接这些网络到因特网是一项挑战。对于“物联网”的成功，标准化是至关重要的。6LoWPAN同时是连接不同物理和链路层协议的一个标准协议和一种融合解决方案，支持将LoWPAN设备连接到因特网。对于“物联网”的成功，至关重要的是设计和部署标准的解决方案，支持基本操作，如路由、移动性以及节点和服务发现与宣告。

致谢

这项工作部分地得到葡萄牙下一代网络和应用组“Instituto de Telecomunicações”

的资助以及“Fundaçõopara aCiência e a Tecnologia”（FCT）国家基金 PEst- OE/EEI/LA0008/2011 项目的支持。

参考文献

1. I. F. Akyildiz, W. Su, Y. Sankarasubramaniam, and E. Cayirci, "Wireless sensor networks: a survey," *Computer Networks*, **38(4)**, 393–422, 2002.
2. IEEE Computer Society, "802.15.4: wireless medium access control (MAC) and physical layer (PHY) specifications for low-rate wireless personal area networks (WPANs)," 2003.
3. G. Montenegro, N. Kushalnagar, J. Hui, and D. Culler, "RFC 4944 transmission of IPv6 packets over IEEE 802.15.4 networks," 2007.
4. N. Gershenfeld, R. Krikorian, and D. Cohen, "The Internet of things," *Scientific American*, **291(4)**, 76–81, 2004.
5. Commission of the European Communities, "Internet of things—an action plan for Europe," Communication from the commission to the European Parliament, the Council, the European Economic and Social Committee, and the Committee of the Regions web page, January 2010, http://ec.europa.eu/information_society/policy/rfid/documents/commiot2009.pdf.
6. J. Hui and D. Culler, "Extending IP to low-power, wireless personal area networks," *IEEE Internet Computing*, **12(4)**, 37–45, 2008.
7. A. Dunkels and J. Vasseur, "IP for smart objects alliance," Internet protocol for smart objects (IPSO) alliance white paper No. 2, IPSO, September 2008.
8. N. Al-Karaki, "Analysis of routing security-energy trade-offs in wireless sensor networks," *Internet Journal of Secure Network*, **1(4)**, 634–660, 2006.
9. A. Chowdhury, H. Ikram, M. Cha, H. Redwan, H. Shams, M. Kim, and S. Yoo, "Route-over vs mesh-under routing in 6LoWPAN," *Proceedings of the 2009 International Conference on Wireless Communications and Mobile Computing: Connecting the World, Wirelessly*, 1208–1212, 2008.
10. T. Narten, E. Nordmark, and W. Simpson, "RFC 1970 neighbor discovery for IP version 6 (IPv6)," 1996.
11. Z. Shelby, S. Chakrabarti, and E. Nordmark, "Neighbor discovery optimization for low-power and lossy networks (draft-ietf-6lowpan-nd-17)," 2010.

12. J. Yick, B. Mukherjee, and D. Ghosal, "Wireless sensor network survey," *Computer Networks*, **52(12)**, 2292–2330, August 2008.

13. A. Reddy, P. Kumar, D. Janakiram, and G. Kumar, "Wireless sensor network operating systems: a survey," *International Journal of Sensor Networks*, **5(4)**, 236–255, 2009.

14. N. Kushalnagar, G. Montenegro, and C. Schumacher, "RFC 4919 IPv6 over low-power wireless personal area networks (6lowpans): overview, assumptions, problem statement, and goals," 2007.

15. J. Al-Karaki and A. Kamal, "Routing techniques in wireless sensor networks: a survey," *IEEE Wireless Communications*, **11(6)**, 6–28, 2004.

16. L. Oliveira, A. Sousa, and J. J. P. C. Rodrigues, "Routing and mobility approaches in IPv6 over LoWPAN mesh networks," *International Journal of Communication Systems*, Wiley, ISSN: 1074-5351, doi: 10.1002/dac.1228 (in press).

17. Ed. T. Winter, *et al.*, "RPL: IPv6 routing protocol for low power and lossy networks (draft-ietf-roll-rpl-19), 2011.

18. T. Narten, E. Nordmark, W. Simpson, and H. Soliman, "RFC 4861 neighbor discovery for IP version 6 (IPv6)," 2007.

19. S. Thomson, T. Narten, and T. Jinmei, "RFC 4862 IPv6 stateless address autoconfiguration," 2007.

20. http://www.tinyos.net/

21. http://contiki-os.blogspot.com/

22. TinyOS, "Blip tutorial," August 2010, http://docs.tinyos.net/index.php/BLIP Tutorial.

第10章　光纤上的IP

Nuno M. Garcia，Nuno C. Garcia

本章描述与普遍成为物理方式上的IP有关的问题，即在一种物理媒介上传输一条IP分组（或数据报）。本章开始时，给出IP的常见协议栈，并从一个比较一般的角度，给出开放系统互联模型，之后介绍数据分组成帧的概念、其动机和关联的功能，接下来是当前光网络（实现波分复用）架构和控制的分析。因为实现这个概念的光网络最适合于核心和城域拓扑（其中使用到数据汇聚），所以本章也从一个语境无关的观点讨论数据汇聚的概念，介绍了一种IP分组汇聚和转换器的机器。最后，本章讨论了一个全IP光网络的一种可能架构，是使用光突发交换范型实现的。

10.1　引言

本章给出物理方式上的因特网协议（IP）的概念，特别讲解波分复用（WDM）上（传输）IP的概念。

研究在物理方式上使用IP，存在明显的动机。如下：

1）IP网络是广泛被接受的，并提供在泛在因特网中的最强大的融合技术，支持不同机器和不同应用成功地交换数据。

2）在一条IP分组上实施的操作越少（从该分组被创建时开始直到被交付到目的主机为止），则它传输得就越快。由此，就需要去除目前在网络中实施的一些冗余操作。

另外，逐渐增多部署的WDM网络，提供传输IP流量的一种可扩展物理媒介，原因是这项技术使用已安装的光纤基础设施，支持因特网流量的增长，并支持网络的带外管理，这将IP传输层与管理和控制层隔离开来[1]。

本章是以这样一种方式组织的，以便能够引入逐步变得复杂的概念，在结束时给出一个全IP光网络的一项建议。这个简短的引言，给出物理方式之上IP考虑光网络的核心动机，之后，本章后面部分组织如下：10.2节讨论封装的概念，并给出开放系统互联（OSI）模型和传输控制协议/因特网协议（TCP/IP）栈；10.3节讨论需要对将被发送的数据成帧，并讨论当前成帧标准，例如以太网和同步数字系列（SDH）；10.4节给出IP和光网络的语境，特别是一些光网络架构，其中包括WDM上的IP（传输）；10.5节讨论WDM网络的控制以及它是如何实现的；10.6节从格式无关的角度介绍数据分组汇聚的概念，并给出IP分组汇聚器和转换

器（IP-PAC）概念；最后，10.7 节给出全 IP 光突发交换（OBS）网络的概念。小结和参考文献结束了本章的讨论。

10.2 封装中的网络数据

在一个网络上两台计算机之间的通信，已经以许多种方式加以分类，包括（依据通信在其中进行的媒介不同）数据被封装的类型、通信的物理和/或虚拟拓扑和通信本身的类型。但通信过程本身的系统化仍遵循两种方式之一，即 OSI 模型[2] 和 TCP/IP 栈[3]。

作为对 Vinton Cerf 和 Robert Kahn 工作的响应，国际标准化组织（ISO）开发了 OSI 模型，Vinton Cerf 和 Robert Kahn 于 1974 年发布了高级研究计划署网络（ARPANET）的 TCP/IP 族的架构。OSI 模型 7 层中的多数层可被映射到 TCP/IP 族的一些层，如图 10.1 所示。图 10.1 的左侧给出了 OSI 模型的 7 层。高层比较接近用户（如使用一个因特网浏览器的一名用户），这与低层相反，后者比较接近或是系统硬件的组成部分。



图 10.1 OSI 模型的 7 层和 TCP/IP 栈之间的对应关系

由用户（一个人或一项应用）产生的数据，将被准备传输到目的机器。这个过程遵循许多步骤，包括将用户数据封装到一个数字信封中，该信封包含所有需要的信息，这些信息支持从源机器和应用到目的机器和应用的正确的数据通信。

用户数据的封装是通过添加后续的数字信封完成的。为了更好地说明这一点，考虑这样一个例子，该例支持将在本章后面做出的一些声明。假定在如下一个地方休假，没有任何类型的计算机或电话，通信的唯一方式是常规邮件，记录一条消息的唯一方式是老旧的纸张和铅笔。您需要将一次会议提醒您的同事，但忘记了将这次会议添加到他/她的日程。您可像这样做：在一小张纸上写下消息，并将这张纸通过常规邮件的方式发送到您的公司，请秘书将这张纸放到这样一个位置，即当您的同事到达时，他/她可以看到这张纸。在信封内的消息可能看起来像图 10.2 所示

的情况。

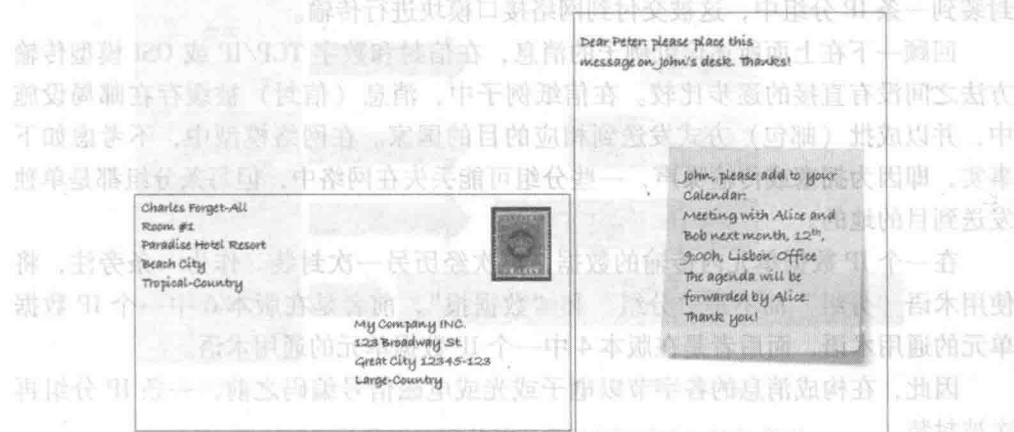


图 10.2 在一张纸上一条消息及相应信封的例子，这说明封装（方法）

当然，这张纸也需要放在一个信封内，而且明显地您将需要在发送者区写上您的地址，在接收者区写上贵公司的地址，不要忘记向邮局支付邮票。这可能会花费一段时间，但最可能的情况是，消息将被交付到您的同事。

在这个样本中，用户（您）得到的是，发送了许多消息。首先，发给您同事的消息将会议提醒了他/她；之后，发给办公室秘书的消息，他首先接收地址为公司的所有邮件；最后，发给邮局的一条消息，说明信件要交付到哪里。这些消息的每条消息都写在纸的一个特定部分，即为了正确地交付消息，不要将办公室的地址写在信封内的纸上，也不要将写给秘书的消息写在信封上。另外，邮局也将使用信封上的消息（目的地址）确定要将信件放到哪个邮包，最后是邮包以合适的飞机或船舶进行运输的。

这个例子的主要目标是说明，需要将正确的信息放在正确的位置，这些消息是为一个特定代理所用的，他将消息转发到一个中间或最终目的地。

在这个例子中，有地址［“123 Broadway St.”（百老汇大街 123 号）］，有如何处理消息［“Please place this on John's desk”（请将这张纸放到约翰的桌上）］指令（消息），有支持操作本身［“Meeting with Alice”（与爱丽丝会面）］目标的信息（在一条消息中）。

再来讨论数字信封。在一个计算机网络中，消息是跨网络从一台源计算机（或网络机器，也称作一台网络主机，或简单称作一台主机）传输到一台目的主机的。

消息是在 OSI 模型（或 TCP/IP 模型）的应用层中产生的，被格式化，并被封装到一般称作协议数据单元（PDU）之中。例如，在一名用户正访问一个网页［使用超文本传输协议（HTTP）传输的一个超文本标记语言（HTML）文件］的情

形中，该网页可能被分段，每部分被封装在一个TCP分段中，TCP分段接下来被封装到一条IP分组中，这被交付到网络接口模块进行传输。

回顾一下在上面所述信纸例子的消息，在信封和数字TCP/IP或OSI模型传输方法之间没有直接的逐步比较。在信纸例子中，消息（信封）被缓存在邮局设施中，并以成批（邮包）方式发送到相应的目的国家。在网络模型中，不考虑如下事实，即因为拥塞或传输噪声，一些分组可能丢失在网络中，但每条分组都是单独发送到目的地的。

在一个IP数据单元内传输的数据，再次经历另一次封装。作为一条旁注，将使用术语“分组”而不是“分组”和“数据报”，前者是在版本6中一个IP数据单元的通用术语，而后者是在版本4中一个IP数据单元的通用术语。

因此，在构成消息的各字节以电子或光或电磁信号编码之前，一条IP分组再次被封装。

下面近距离地看看在欧洲一个典型家庭中的一台计算机和在美国某处一个内容提供商之间的一次通信中发生了什么。在这个例子中，该用户尝试访问存储联合国主页的一台服务器。为做到例子的清晰，假定用户的家庭有一个小型局域网，它由一台或两台计算机、一台交换机和一台不对称数字用户线（ADSL）调制解调器组成。ADSL是一项异步技术，它支持一个因特网服务提供商（ISP）销售铜线上的宽带因特网接入［高达每秒数兆比特（Mbit/s）］。数据将一定使用几条洲际大西洋光纤线缆之一，并最终被交付到在纽约为联合国提供接入的ISP。

首先，用户在浏览器的地址栏中输入一个地址。例如，用户输入http://www.un.org，这是联合国网站的官方地址。在联合国的HTTP服务器可对请求这个文件的主机（家庭中的计算机）做出应答之前，涉及许多协议［通常描述为域名系统（DNS）协议］，它们将名字www.un.org转换为一个相应的IP地址。在这种情形中，与这个名字关联的因特网协议版本4（IPv4）[4]地址是157.150.34.32。

之后，用户的应用将一条请求发送到HTTP服务，这是运行在作为服务器的机器（或多台机器，也许是实施其他任务的几台机器，如负载管理、防火墙等）中的一个程序。依据HTTP，格式化这条请求，并将其封装在一个TCP分段内。这个TCP分段被封装在一条IP分组内，之后这个分组被交付到网络接口层进行传输。图10.3给出了一种方案，说明通常称作封装的概念。

在用户的计算机中实施的封装，产生一个以太网帧，该帧将被传输到用户的路由器或路由器调制解调器。这里，调制解调器将去除以太网首部和帧校验序列（FCS）（是在用户的机器处由网络接口层产生的），并添加另一个首部（和可能的另一个FCS），以便将IP分组从用户的家庭转发到ISP的设施。路由器也许将用户机器放置的IP源地址改变为路由器的公开IP地址，这是称作网络地址转换（NAT）的技术。

一旦分组进入ISP线缆和网络设备，就只能猜测分组所经历的变换了。在OSI

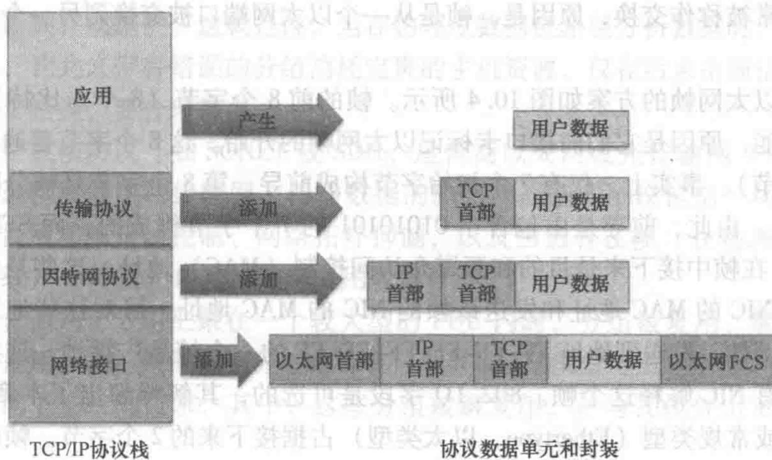


图 10.3　说明 TCP/IP 栈的方案和在每层产生的各 PDU

模型的低层上存在许多协议，因此，可使用许多协议和格式来传输该分组，或再次使用邮局的比喻，ISP 可为用户的原始消息添加几种信封形状和颜色。

在大西洋洲际光缆之一中穿越时，分组也许使用 SDH 或同步光网络（SONET）协议进行传输。

为使一个冗长的例子变得简短，在用户的计算机中创建的 IP 分组在它到达地址为 157.150.34.32 的目的地之前，几乎保持不变，例外情况是由它在传输路途中可能遇到的 NAT 设备做出的改变。但是，每次分组从一个网络传输到下一个网络时，即从用户的网络到 ISP 网络，之后从 ISP 网络到运营大西洋光缆的 ISP 等，分组都将被重新封装到一个新的信封（首部）中。

实施 IP 分组的封装，支持网络实施一些基本任务，例如将数据交付到正确的机器，实现的功能确保网络传输是灵活有弹性的，甚至在某个点支持对客户的计费。

10.3　为什么需要帧

要判定成帧何时首次被用来封装数据是困难的，但追溯到 20 世纪 70 年代，原始的以太网协议已经为净荷考虑一个成帧结构。以太网协议的主要目标是通过一个常见媒介（如同轴线缆）传输数据分组。因为最初的以太网传输媒介是以一种广播方式工作的，所以冲突的检测是由网络接口卡（NIC）实施的，NIC 检测线路上信号功率的增加，这使之能够中断帧的传输，并在后来的时间重试帧的传输。这种方法被称作带有冲突检测的载波侦听多路访问（CSMA/CD）。

在一个局域网（LAN）上，在一个以太网帧内部一条 IP 分组的封装几乎排他性地使网络设备将一条分组从一台机器交付到那个网络内的另一台机器，这样的一

项任务通常被称作交换，原因是，帧是从一个以太网端口被交换到另一个以太网端口的。

一个以太网帧的方案如图 10.4 所示。帧的前 8 个字节（8 个 8 比特）具有一项同步功能，原因是它们向接口卡标记以太网帧的开始。这 8 个字节普遍被分类为前导（字节）。事实上，仅有 7 个初始字节构成前导，第 8 个字节是帧分隔符的开始（SFD）。由此，前导是由包含字 01010101 的 7 个字节构成的，而 SFD 包含字 0101011。在帧中接下来是目的和源媒介访问控制（MAC）地址。这些是假定要接收该帧的 NIC 的 MAC 地址和发送该帧的 NIC 的 MAC 地址。可对这些地址施用一些惯例。例如，发送到地址 FF-FF-FF-FF-FF-FF 的一个帧是广播帧，连接到那个媒介的所有 NIC 解释这个帧。802.1Q 字段是可选的，其解释超出了本章的范围。帧的长度或常规类型（Ethertype，以太类型）占据接下来的 2 个字节。帧以 FCS 结束，这是一个 32 比特循环冗余校验（CRC），支持错误检测。以太网也规范了 12 个字节的一个帧间隙。净荷由最小 46 字节和最大 1500 字节组成，由此通常确定一个局网的最大传输单元（MTU）是多大。这就是广泛被认为 99.99% 以上的因特网流量由 IP 分组组成的原因，这些 IP 分组多数都是 1500 字节长的[5]。

前导	SFD	目的 MAC	源 MAC	802.1Q (可选)	以太类型或长度	净荷	FCS	帧间隙
7	1	6	6	4	2	46~1500	4	12

图 10.4　一个通用以太网帧的方案，在顶部给出帧的各组成，在底部给出它的期望长度（以字节表示）

由帧结构的分析，可推断出与成帧规程相关联的一些功能。如下：

1）使接口卡可检测一个帧的开始序列。

2）支持错误检测。

3）使媒介能够保持在一个静默水平，其中 NIC 可检测没有发生通信的情况。

这三项功能可被容易地中继到物理层，原因是这些是可由 NIC 实施的真正任务。最近成帧协议（如 SONET 或 SDH）不考虑前导（字节），后面介绍这些协议。

在局域网和接入网中情况是这样的，但在更大范围的网络上，IP 封装也支持分组在传输网络内交换，并支持实现许多其他的功能特征。

包含 IP 分组的帧以及使 IP 分组在一种物理方式（如光纤或铜线，或使用无线电磁波，像在无线传输中的情形）上传输的帧，被用来支持网络实施许多重要的任务。

可生存性、可用性和控制是这些任务中最重要的任务，对网络运营商而言是至关重要的问题，他们期望网络基础设施遵循 5 个 9 的规则（即成功率必须至少为 99.999%），成功地交付数据。应该在通信的最低层次，实施跟踪和纠正错误的任务，这意味着越早检测到错误，则错误被纠正得就越快，或网络可请求带有纠正过

的数据的新帧，就越快。这就是说，当在物理或数据链路层分析数据时，检测一次传输错误，比允许带有错误的分组消耗宝贵的主机资源，仅在后来由通信栈的高层检测到这个错误，要合理得多。

这就是成帧协议［如SONET或SDH、运营商以太网或光传输网（OTN）］普遍为电信运营商使用在传输网上传输数据的原因[6]。这些协议包括一项开销，使网络运营商可实施错误控制、网络拓扑抑制，以及当然有在帧［在物理媒介（如光纤或导线）中传输］内复用和解复用数据分组。

据称，当几个分组汇聚在一个较大型的PDU内时，分组被复用。解汇聚任务被称作解复用。被汇聚的数据分组也许有一些共同的性质，也许它们具有被传输到网络中相同下一跳的需求，其中，这些分组被解复用，并与其他分组重新复用在一起。

SONET和SDH协议是定义在准数字系列（PDH）协议上面的，PDH要追溯到20世纪60年代中期。SONET大多用在北美，而SDH大多用在欧洲和日本以及多条洲际链路。开发SDH协议，是为了支持话音通信的传输，其焦点是数字话音电路的复用。话音通信使用4kHz的带宽，并使用8比特的采样尺寸以8kHz进行采样（足以表示完整的带宽）。由此，8比特乘以8kHz，得到64kbit·Hz的比特率，即64kbit/s。因为运营商需要传输几个数字话音流，所以定义了这个基本64kbit/s单元的整数倍。对于SONET，基本的64kbit/s被命名为数字信号0（DS0），接下来是整数倍的：DS1 =1.544Mbit/s、DS2 =6.321Mbit/s等，其中DS4 =139.264Mbit/s，即DS4 =4032个用户信道。针对SDH，定义了类似的整数倍速率：第一个整数倍速率，E1 =2.048Mbit/s（或32个用户信道），E2 =8.448Mbit/s（或128个用户信道）等。

针对DS1，计算如下：每个DS1电路包含24个用户信道，每个这样的信道都使用一个8比特字编码。另外，每个帧需要一个成帧比特，每秒总共有8000个帧，即

（8比特/信道×24个信道/帧+1比特/帧）×8000帧/秒 =1544000比特/秒 =1.544Mbit/s

应该注意到，经常的情况下，当涉及通信单位时，兆（M）意味着10^6，而不是像计算字节时通常所说的1024×1024，其中千（k）指2^{10}而不是10^3。

针对E1各整数倍速率，可进行类似的计算。

SONET和SDH包括首部字段，如分节开销和线路开销，它们包含有关帧结构、错误纠正和自动保护交换消息的信息。图10.5给出了同步传输模块1（STM-1，SDH第一级的数据单位）的一个样例方案。其中，在这个数据单位中的各字节是按顺序逐字节的、逐行地传输的，AU表示管理单元。

这些字段携带的数据，使运营商能够部署网络的性能监测、连接和流量类型的识别、链路故障的识别和报告，以及错误检测和/或纠正。

除了在首部中信息所支持的功能外，封装的额外开销是人们所不期望的，原因是它消耗网络和主机资源，这是因为一方面这意味着必须传输更多字节，另一方面主机（网络中的路由器或机器）需要处理封装帧，以便支持数据分组的检索。

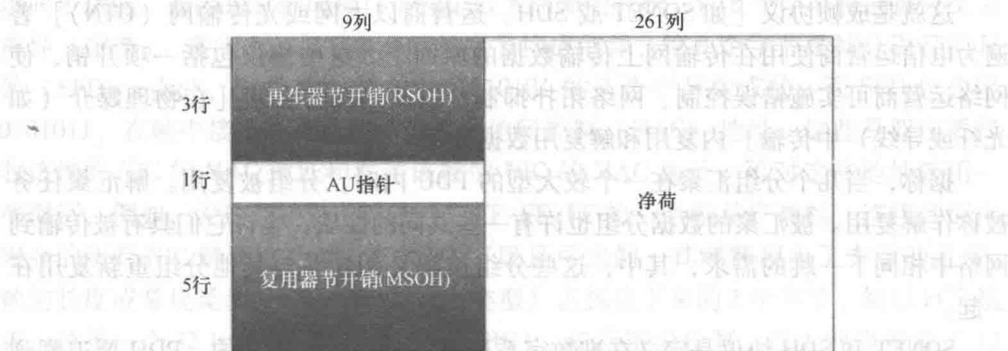


图 10.5 STM-1 数据单元跨越 9 行 270 列的方案

10.4 IP 和光网络

产生对各项技术的一种更高效替代方案（产生另外一个封装级）的需求，导致物理方式上传输 IP 范型的建议，特别是 WDM 上传输 IP 的建议，这是一项复用技术，支持在一根光纤中产生数个数据信道（波长信道）。粗略地说，每个数据信道是由一个激光器（以一个不同的光传输数据）插入到光纤中的，所以一组激光器（每个激光器以脉冲方式产生一个不同的光）产生不同的光脉冲（可在一条光纤中传输）。

在 WDM 中，假定每个波长是能够传输数据的一条数据信道。虽然 WDM 技术仍然较新，多种器件（如波长转换器、光随机访问内存等）仍然在密集研究之下，但目前存在用于实现数据信道复用的几项技术，每个信道以一个特定波长被编码到单条光纤。这也被称作 λ 复用，原因是一条数据信道是在光纤中一个特定 λ（一个特定波长）中传输的。另外，因为光纤非常低的信号衰减性质，所以它是运输信号的一种卓越媒介。

考虑 WDM 上传输 IP 法为核心网最具前景的传输技术，有两个主要原因。第一个原因是处理 OSI 模型的扁平化。通过直接在一种物理媒介上传输 IP，就去除了数据链路层，由此，支持网络执行的功能是由 IP 层假定的（如交换是在网络地址上实施的，像在 OBS 中的情形），或由物理层假定的（如错误纠正和帧排序）。一些功能根本不执行，如一个机器地址及其相应 IP 地址之间的指派。

采用在 WDM 上传输 IP 的第二个原因是 WDM 技术的固有潜在的（大）容量。每个 λ 信道（每个波长）可支持高达 100Gbit/s 的数据速率。商业解决方案目前支

持每条光纤中高达80个波长，虽然2005年日本电报和电话公司报告，在Kyoto Keihanna和Osaka Dojima实验室之间在126km测试床上已经测试过1000个数据信道[7]。存在大量数据信道的情况，其中假定在商业或商用前设备中每个信道以从2.5Gbit/s到100Gbit/s范围的速率传输数据，这支持为核心网基础设施提供大量的带宽。容易得出结论，单条光纤可每秒传输数太比特的数据，这种容量主要受限于端设备及其将多个信道复用到一根光纤和/或以较快速率传输数据的能力，即这不太依赖于光纤本身。当然，较佳工程化的光纤将支持使用新频带的波长。

WDM网络在一个独立的数据信道中传输控制和管理数据，这个信道称作控制或监控信道。控制信道负责传输与网络管理和管理架构有关的数据。

虽然这不特别是本章的主题，但进一步阅读参考文献［1，6，8-12］中有关OBS和光网络的材料，也许会发现是有助益的。

WDM技术上传输IP，主要是通过面向电路的架构实现的，原因正如前面讨论过的，还没有能够执行光交换的光逻辑器件，由此构造点到点的一个全光网络。WDM或光网络，一般而言，仍然依赖于光电光（O-E-O）信号转换，支持以数据单元表示的信息以电形式在每个网络节点加以解释。

比较简单的光电路交换架构，如“光网架构”[13]，显示出作为技术发展方面的一些潜力，这种技术支持一根光纤中数据信道数量的增加。这种方法使用WDM技术，特别是在相同物理链路中几个数据路径（λ）的存在，可增加用户可用的吞吐量，并简化交换。一条光路径（数据信道）是在网络中两个节点（不必是相邻的）之间建立的，是通过在整个路径上分配相同波长产生的。一旦建立光路径，则中间节点就不实施处理、缓冲或电光（E-O）转换。当与常规的存储转发网络相比较时，使用光路径建立电路，并由此传输分组的做法，降低了整体的网络缓冲和处理需求。在这个架构的基础上，得到一个集成的分组和电路交换解决方案，原因在于分组是在邻接光路径上路由的，在电路存在期间，电路是使用路径上可用的数据信道建立的。

图10.6给出了一个4节点环网络的样例光路径需要。光路径网络的高效管理焦点在解决两个问题上：首先，波长是一种稀有资源，由此就所用的波长数而言，高效地建立光路径是必要的；第二，当与一条光路径建立（其中不解决这个波长连续性约束，即如果存在波长转换器并使用波长转换器）相比时，使用整条路径的相同波长建立一条光路径的需求，引入了潜在的带宽浪费。当需要建立一条光路径电路而引入另外的带宽浪费时，则光路径网络操作出现第三个问题。见图10.6，这样的一个例子将是如下一种状况，其中节点2和节点3之间的流量是非常密集的，而一些数据信道为节点1和节点4之间的流量与节点2和节点3之间的流量所预留。在这个例子中可论断，节点2和节点3之间的链路是路径过载的，而节点1和节点4之间的链路仅有一条指派的路径。此外，考虑架构的这些类型的静态或准静态特性，以非常规和不可预测流量的观点看所使用的网络资源，将总是存在低效

情况。

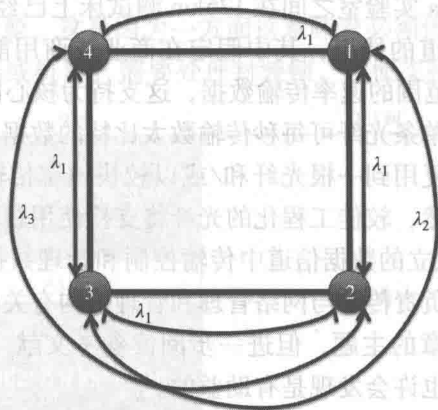


图 10.6　带有 4 节点和双向光路径的光网架构

虽然当前存在的许多光电路架构没有特别将焦点放在物理方式上传输 IP，但它们可演进包括这样一种场景，此时所用的许多光交换机可处理多协议传输，例如在参考文献［14］中描述的路由器。

WDM 上（传输）IP 的做法，有时以仍然需要成帧的一种观点加以对待[15]。数据链路层实施的功能包括：

1）成帧（如以太网）。

2）分组汇聚（如 SDH 或 SONET）。

3）错误检测（如 CRC）。

4）错误恢复［如自动重复请求（ARQ）］。

在前面论证过，即这些功能中的一些功能可由接口本身执行，如错误检测和恢复。如后面将看到的，其他功能可在不需要成帧的条件下执行，即在不需要使用一个帧中信息的条件下，利用 WDM 的带外信令特征，实现汇聚和链路抑制、存活性和控制是可能的。

在 WDM 上传输 IP 的范型，通常被看作是遵循三种可能的方法之一加以实现的。事实上是从 WDM 范型本身派生的方法。如下：

1）电路交换 WDM 网络，如光路径架构。

2）突发交换 WDM 网络，如 OBS 架构。

3）分组交换 WDM 网络。

如前所述，还不存在支持一个全光端到端的分组交换网络的实现技术。这意味着，在每个节点处，每条分组（或每个 SDH、SONET 数据单元）需要考虑电域（通过光电转换），在其中加以解释并路由到合适的接口。

电路交换网络，诸如在光路径架构中建议的网络，考虑到多个光信道的可用性。在这些架构中，电路是预定义在网络中的，中间节点不了解数据的路径或其格

式（如分组或SDH帧）。

这意味着，对于这些网络，SDH成帧主要服务于两个目的，即将支流IP分组汇聚成单个数据结构和支持使用SDH首部中的数据来执行网络监测任务。

OBS网络架构尝试在电路和分组交换（还不可用的）之间实现一个中间的但可行的粒度级。OBS网络考虑到入口节点处的数据汇聚和临时创建一条电路，以支持跨网络的突发，但仍然维持中间节点不知道它的通路。将在下面各节讨论这种范型。

所考虑的WDM网络上传输IP的一个通用架构如图10.7所示[6]。可看到：

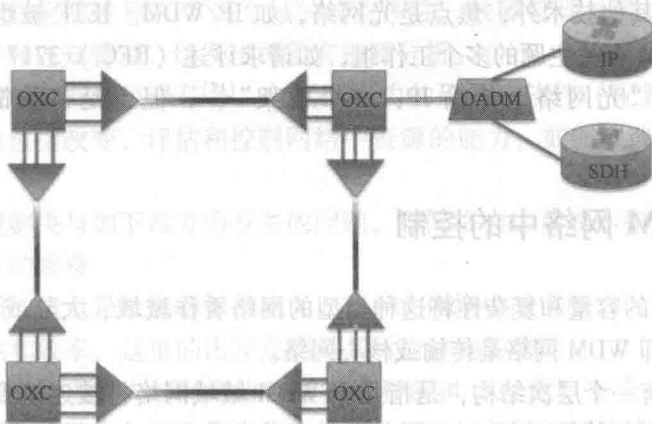


图10.7　在WDM网络上传输IP的一种可能架构的方案，给出相关的组件

1）连接4个WDM节点的彩色线，代表光纤，每条光纤传输多个λ信道。通常至少有两条光纤，每条光纤在一个方向上传输光子。这组光纤通常被称作一条链路。

2）这些光纤连接到复用器/解复用器（这里表示为三角形），它们的功能是，当传输光信号时，将不同波长复用进入一条光纤，并将光信号解复用回独立的波长。

3）每个波长连接到一个光交叉连接器（OXC），该设备的功能是将波长从一条光纤交换到另一条光纤。

4）光上下复用器（OADM）用作一个流量入口和出口节点，将波长添加到OXC或在IP或WDM终端设备［可能是光线路终结器（OLT），它被连接到IP或SDH路由器］处分出波长。

配置比较复杂的拓扑是可能的，包括这样的拓扑，其中OADM作为一个简单环或线性网络中的一个网元（NE）。OXC是这样一台设备，它支持将一个特定端口（一个波长）的信号切换到另一个端口。WDM NE集成复杂功能，如波长转换、功率补偿和信号再生。

在图10.7所示的方案中，连接到IP路由器的一个分支IP网络（图中没有给

出），将找到其分组，由路由器转换为光形式，并被传输到最近的 OADM。之后 OADM 将光信号转换为一个特定的国际电信联盟（ITU）波长，以便使它能够集成它当前所复用的波长集合。一个 ITU 波长是这样一个波长，它被定义为一个信道和在所定义频带之一中的一个频率。例如，一个密集波分复用（DWDM），ITU 定义 C 频带上的信道 44 为 1542.14nm 的波长[16]。

就 IP/WDM 标准化方面，存在许多研究工作。感兴趣共同体有两个主要焦点组，一个是由因特网工程任务组（IETF，www.ietf.org）促成的，另一个是由国际电信联盟标准化部门（ITU-T，www.itu.int）促成的。ITU-T 被组织成研究组。第 15 研究组，除其他技术外，焦点是光网络，如 IP/WDM。IETF 被组织成工作组。存在与 IP/WDM 有关主题的多个工作组，如请求评述（RFC）3717 的工作组，该 RFC 的标题是“光网络上传输 IP：一个框架”[17]，但也是有关链路管理协议的[18]等。

10.5　WDM 网络中的控制

WDM 网络的容量和复杂性将这种类型的网络看作城域、大陆或洲际数据传输的理想网络，即 WDM 网络是传输或核心网络。

核心网络有一个层次结构，是指几个 WDM 城域网络将被连接到国家级 WDM 网络，这接下来与其他国家级 WDM 网络一起将连接到一个大陆级网络，这个网络接下来连接到全球核心网络。图 10.8 给出了这个情况的图示。在传输和核心级采用 WDM 网络的原因是 WDM 网络提供快速的和可靠的光数据传输。但这带来保障不同 WDM 网络互联的问题。

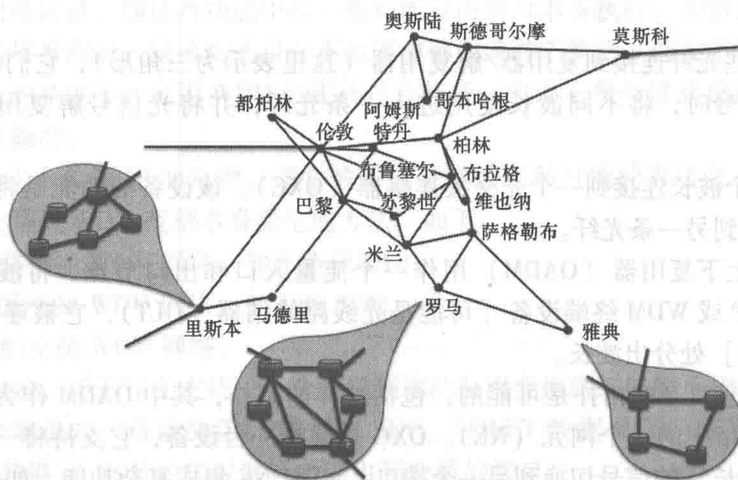


图 10.8　在 EON（欧洲光网络）的拓扑上（由 19 个节点组成的 EON 拓扑）给出不同层次 WDM 网络的方案，带有到里斯本、伦敦和莫斯科的其他 WDM 网络的可能连接

在物理层，通过应用如下设备做到信号兼容，这些设备如可重新配置的光上下复用器（ROADM），可提供光信道（或 λ）在局部网络汇聚点的插入和摘除。这些设备也许将需要使用波长转换器，将一个给定的到达波长转换到当前存在于输出光纤中的另一个波长。

在管理层，每个网络由其相应的权威机构管理，并可能集成一个或几个自治系统。为 WDM 网络所采用的网络控制和管理（NC&M）框架通常是著名的 OSI 管理框架 FCAPS，带有 5 个功能域：故障管理、配置管理、计费管理、性能管理、安全管理。

现在将简短地描述 FCAPS 每个域的焦点所在。故障管理处理故障检测、隔离和纠正，告警和通知管理，以及也可能有根源分析。

配置管理负责资源目录任务，包括网络内的硬件、软件（配置）和电路。资源目录任务也包括改变、评估和控制网络中资源的能力，如能够改变一个给定 NE 的配置。

计费管理解决与如下两方面有关的问题，即网络的运营成本和为占用网络资源的用户所建立的缴费。

性能管理的焦点是性能指标的定义、指标在网络中的实施，以及将之用于评估网络的有效性和效率，这里的出发点则是其所应达到的功能。

最后，安全管理解决与网络的物理安全和逻辑（运营）安全有关的问题，这里的安全指网络的状态，其中所有前面提到的管理域都是以可信赖的方式由网络运营商可部署的。另外，对于光网络，安全层也处理运营安全问题，原因在于包含激光的各 NE 要符合激光物理安全规范是强制性的要求。

FCAPS 的许多域不在本书范围，将不做讨论。但在 WDM 上传输 IP 的语境中，不得不在 WDM 网络管理问题中稍稍挖掘得深一些。

前面介绍过的 ITU-T 研究组，开发和提出一个框架，该框架支持动态电信服务的部署、管理和运营，称作电信管理网（TMN）。TMN 支持不同的和异构的网络通过一种面向对象的方法进行互联和通信，这种方法定义了许多对象。电信网络本身被定义为一组设备（交换系统、电路、终端），一般称作 NE，由运营支撑系统（OSS）监测，OSS 是支持 NE 运行和监测的设备和软件，即 OSS 支持在各 NE 上实现 FCAPS 功能域。

TMN 模型也以一种层次方式加以实现，每层处理与 FCAPS 模型有关的问题。TMN 模型的 5 层是商务管理层、服务管理层、网络管理层、网元管理层和网元层。顶部各层解决诸如商务级和服务级协议和约束的问题［包括与服务质量（QoS）有关的问题］，底部各层处理诸如拓扑、NE 配置以及错误和状态消息等问题。

TMN 模型使用称作 Q3 的一个消息接口，将消息从网络本身传递到网络管理操作员。图 10.9 给出了一个 WDM 网络及其关联的 TMN 架构。在这个图中，可看出 TMN 服务是如何由一个服务提供商提供的，由此使客户/网络拥有者在开发商务域

方面获得自由。

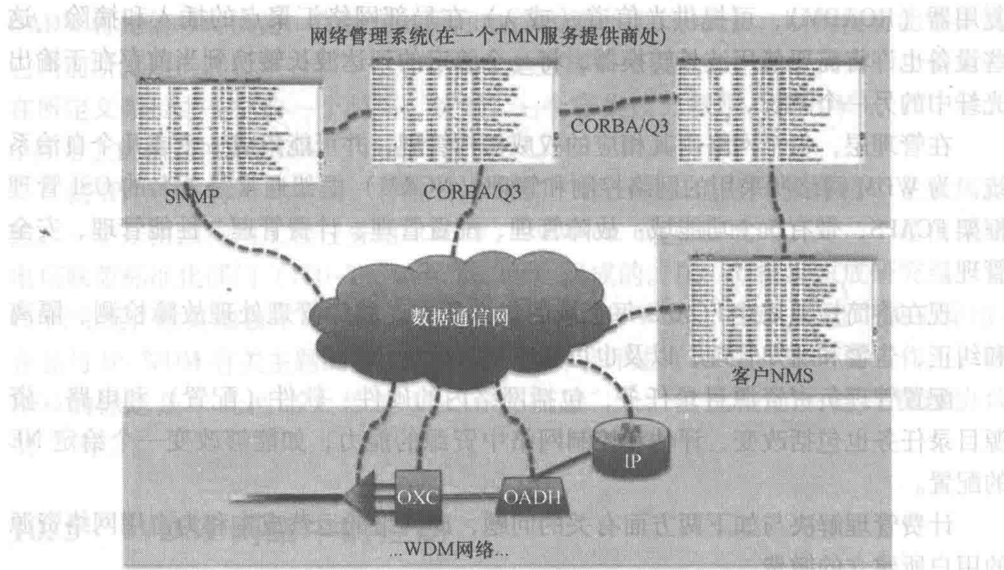


图 10.9　一个 TMN 架构及其关联的 WDM（部分）网络的方案

存在网络管理的其他协议。也许部署最多的是基于简单网络管理协议（SNMP）的，这是由 RFC 1157[19]定义的一个基于 IP 的协议。与 SNMP 相关联的信息模型被称作管理信息库（MIB）[20]。

TMN 模型定义通用管理信息协议（CMIP）为在网络部件和网络管理系统（NMS）的软件代理之间交换消息所用的协议。基于 TMN 的其他管理模型也许会集成基于 SNMP 或通用对象请求代理架构（CORBA）[21]的消息接口。

在 WDM 网络上传输 IP 的情形中，SNMP、CMIP 或 CORBA 消息可由光监控信道（OSC）来传输，或在 OBS 网络的情形中，NMS 消息可由控制信道来传输。在任一情形中，控制或监控信道都是一条单独的光信道，在每个活跃的网络节点处都遇到 O-E-O 转换和解释，这使每个节点能够在信道中插入一条消息（目的地为运营商的 TMN 网关）。

在图 10.9 中给出的数据通信网（DCN）有一些特殊的特点，原因是即使当在 WDM 网络（假定要传输该网络的管理数据）中存在一次故障，该网络也需要是可运行（操作）的。这通常意味着，DCN 必须有带有一定冗余度的一个配置（拓扑），这样的冗余度确保出现一次光纤故障时 DCN 的正常运行。通常 DCN 是以如下方式实现的：通过光层外部的一个独立带外网络，如使用专用租赁线路。如前所述，通过 OSC，这是最常见的选项，但要求设备（是不了解内容的）的附加逻辑，如用于 OXC 的信令引擎（SE）（后面讨论），或使用开销技术，这种技术是在速率保留的带内光信道上实现的。

10.6 IP域中的分组汇聚

将分组汇聚到突发中的做法，是在20世纪80年代由Haselton[22]和Amstutz[23,24]独立提出的一个概念。设计这种交换方案，是为了从统计复用效果中受益，当共享相同目的地的几条分组被放在一起发送，由此仅需一个首部和一个成帧间隙，此时发生这种统计复用。在20世纪末，这个概念导致提出OBS[8,25-28]，指被归组的各分组的首部，现在是在网络的一条专用信道中传输的一条控制分组，被归组的分组被看作一个格式无关的无首部突发，该突发注定以一种全光形式穿越光网络。

受到OBS研究方面的驱动，在学术界和工业界，焦点也转向突发交换研究。但是，对于OBS范型而言，突发组装算法并不是独特的，由此可由其他传输或交换方案自由地加以采用。

采用人工流量[29-35]，人们研究了突发组装算法的性能，这是指突发组装算法的效率方面以及所提供QoS和突发特点之间的关系、突发流量统计和突发流量刻画等方面，进行的性能研究。对于真实的IPv4流量，在参考文献［36-37］中发布了比较性的性能评估。

数据分组汇聚或突发组装是这样一个过程，其中在得到数据聚集体（也称作突发）之前，个体数据实体［如IP分组、以太网帧、异步传递模式（ATM）信元等］被归组在一起，将突发发送进入网络。突发可能被重新封装（或不会），这取决于所支持的网络场景。以一个附加的信封（首部）封装突发的做法，将路由和处理能力添加到这个大型分组，其净荷是突发，例如突发由一个IPv6分组或巨型数据报的净荷组成的情形。如果突发没有被重新封装，那么它必须以一种透明的方式传输到网络，原因是在字节的聚集体中没有可用的路由信息。

突发的各组成数据分组的特点和源发点与突发组装原则是无关的。突发交换概念仅要求传输信道的另一端运行一个临时的突发解组装过程，该过程检索原始的组成分组，以便将它们进一步路由到目的地子网络。

突发组装算法是约束驱动的，并被分为三类：

1）最大突发尺寸（MBS）[39]。
2）最大时间延迟（MTD）[40]。
3）混合组装（HA）[41,42]。

其他突发组装算法，如考虑服务类的算法，典型地是构建在前述基本类型上的。

在突发组装中的基本原理是，一个突发组装队列收集并管理具有一个共同目的地和一个共同QoS约束集的各分组。遵循这个原理，目的地为一个给定源且是低优先级的各分组，与具有相同目的地但被标记为高优先级的那些分组不同，被组装

在一个不同的队列中，例如，被标记为电子邮件或新闻内容要比TCP或实时协议(RTP)（要了解IP分组中的一个完整协议集，请参见参考文献［43］）具有较高的时间阈值的那些分组。

目前，IP正在进行从版本4到版本6的缓慢迁移。这导致这样一种状况，其中一些子网络仅采用IPv4工作，其他一些子网络采用IPv6工作，而另外其他的子网络具有处理两种格式的能力。自然地，IPv4与IPv6原生网络的通信，意味着使用一种设备，它以如下一种方式将发送方原生格式转换为接收方原生格式，即得到的数据流可被正确地加以解释。

随着IPv6原生系统的份额在市场上的增加[44]，对改善这些系统与遗留IPv4子网络（在已部署的系统中仍然构成大多数）的相互通信，存在日渐增长的需求。人们提出了许多解决方案——隧道、重新封装、变换和转换[45-54]，其中一些方案可以一种积累的方式加以使用。

但没有哪种现有解决方案，可处理IPv6数据报格式所增加的分组容量的有效利用问题[55]，在功能方面而言是有限的。

在本节中给出的机器概念，即IP-PAC，是针对IP分组汇聚器和转换器[56]的，目的在于改善IPv4和IPv6分组在IPv6或IPvFuture（IP未来版本）路由结构上传输的封装和传输效率。在一般情形中，这个机器概念可被看作一种通用的IPv4/IPv6到IPv6/IPvFuture转换和汇聚机器。

IP-PAC算法提出如下动作的一种组合法：

1）遵循一组规则和约束，汇聚IPv4/IPv6分组。

2）如有必要，将以前汇聚的分组聚集体重新封装到一个IPv6/IPvFuture(分组)。

使用上述两个过程的组合，不管分组属于哪个流，一系列到达IPv4/IPv6分组可被组合成单条IPv6/IPvFuture㊀分组。分组汇聚和重新封装过程是在逐条流的基础上实施的，指个体数据流从一条更一般的到达数据流分离。一旦分离，个体数据流就会遇到汇聚，并最终重新封装为一种传输分组格式㊁（IPv6/IPvFuture)。

遵循HA算法，所应用的汇聚方法有两种：

1）时间跨度是动态地在第一条分组和最后一条分组之间设置的，这些分组被汇聚为单条传输分组。这个时间跨度定义一个平均分组时延，所有到达分组都受到这个约束。此外，为汇聚时长设置的时间跨度值，取决于网络负载和到达数据流的自相似度。

2）最大尺寸是针对传输分组动态设置的，这取决于当前网络状况、链路负载和数据流自相似度。在IPv6分组的情形中，最大可接受尺寸是4GB（巨型选项[38]）。

㊀ 从此时开始，将仅说IPv4到IPv6，指IPv4/IPv6到IPv6/IPvFuture变换/封装/汇聚。

㊁ 一条传输分组指这样的一条IPv6/IPvFuture分组，它由汇聚和重新封装机制产生，并通过核心传输网传递到拆装点。

机器概念

因为 IP-PAC 是一个机器概念，它可被用在如下网络中，其中数据传输可受益于突发组装过程所带来的统计复用增益。在这种意义上，IP-PAC 的工作环境会是异构的，如图 10.10 所示。这里给出几个子网所连的一个核心网络（一些子网有一种原生 IPv4 格式，其他一些子网有路由器和/或网关）和直接连接到核心网络的一台路由器。IP 流汇聚是由 Schlüter 于2000 年解决处理过的[29]。从主机到主机到源到目的流的降低因子，被证明主要取决于主机到主机的流平均数［开始于一个区域 s（源），目的地为一个区域 d（目的地）］。Schlüter 提出，取得高降低流比率的最佳环境是一个骨干路由域。

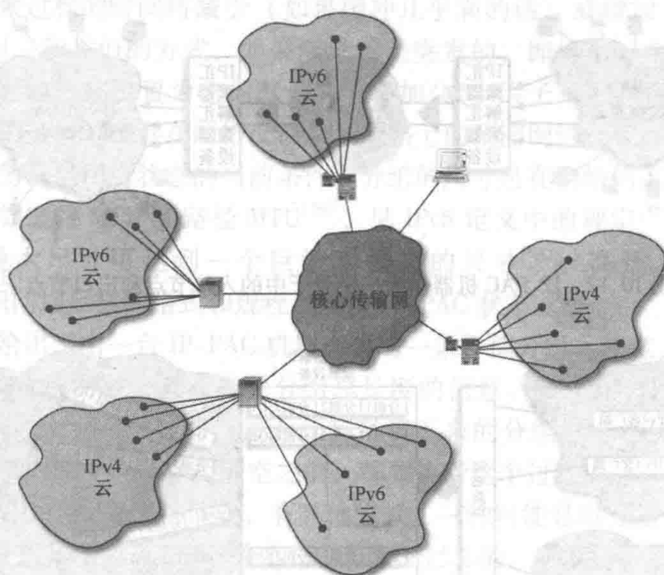


图 10.10　在一个传输网络（带有非 IP-PAC 机器）的边缘处，IP-PAC 机器的样例使用情况[1]

由源 IP-PAC 机器产生的分组具有如图 10.11 所示样本分组的一个格式，例外是如下情况，其中不应该实施一个 IPvFuture 首部的汇聚和/或添加：

IPv6首部(+扩展)	IPv6分组(首部+净荷)	IPv4分组(首部+净荷)	IPv4分组(首部+净荷)	IPv6分组(首部+净荷)	…	IPv6分组(首部+净荷)
首部	净荷					

图 10.11　一条 IP-PAC 产生的分组方案[1]

1）如果净荷仅由一条分组组成，且这条分组已经是转换过的格式（IPv6 或 IPvFuture），即已经是核心网中所用的格式。

2）如果目的地没有分组拆装能力，例如当已知目的地是一台非 IP-PAC 机器时的情况。

3）如果传输网络没有从具有另外一个首部得到好处（即由该首部可做出路由决策），即在一个 OBS 网络的情形中。

在后一种情形中，IP-PAC 应该仍然要对外发分组排队，以便将之按顺序发送，由此仍然从核心网络交换的最小化付出（由路径调节现象[57]带来的）受益。

为支持 IP-PAC 机器将一条 IP 分组识别为一条汇聚分组，汇聚分组的首部（如果存在的话）应该将流量类字段的 3 个高位比特加以设置（见参考文献[55]）。我们称这个过程为 IP-PAC 打标志，称一条分组为“IP-PAC 打过标志”，如果在首部的流量类字段中它的 3 个高位比特进行了设置。

对 IP-PAC 和非 IP-PAC 机器之间如何通信的一个简短解释，可建立如下（见图 10.10、图 10.12 和图 10.13）。

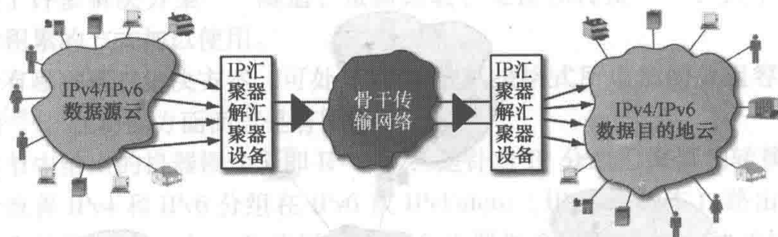


图 10.12　IP-PAC 机器作为一个骨干中的入口节点和出口节点[1]

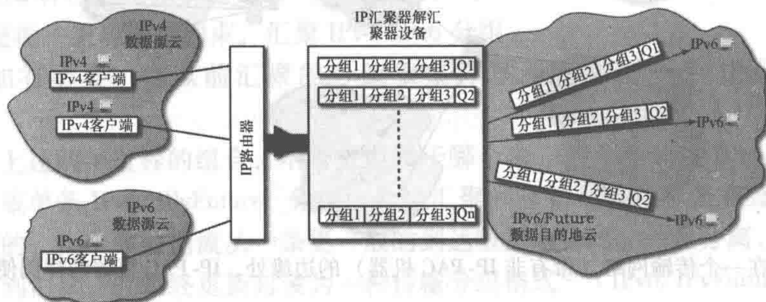


图 10.13　IP-PAC 作为一台网关[1]

如果目的地是一台非 IP-PAC 机器，如一台服务器、一台路由器或一台网关，它们极可能没有拆装能力，则源分组（非汇聚的，但仍然是排队的）以最小时延按序从 IP-PAC 机器发送。

如果分组的源是一台非 IP-PAC 机器，但目的地是一台 IP-PAC 机器，那么在接收到一条分组时，IP-PAC 将尝试寻找 IP-PAC 标志。在这种情形中，因为分组是从没有汇聚能力的一台机器发送的，所以分组被直接解释并转发到相应的出接口（在查询一个内部路由地址表之后做出这项操作）。

图 10.13 给出了一台 IP-PAC 机器如何实施突发组装过程，并将得到的突发

(在一个IPv6/IPvFuture信封中带有封装或不带封装）转发到一个目的地网络。

如果一条汇聚分组的目的地为另一台IP-PAC机器（见图10.12)，则源机器将传输类似于如图10.11所示分组的一条分组。所传输的分组是打过IP-PAC标志的，例外是在本节开始部分之前提到的条件之一，即它们不会遇到如下情况，即IP-PAC发送方机器添加一个额外首部。在这种情形中，分组的目的地址将是目的地IP-PAC机器的地址。

因为HA可模仿MBS和MTD，所以考虑将其作为可应用的突发组装算法，遵循参考文献［37］中的结论，它的突发组装阈值将是动态可调整的。通过使用到达流量的自相似度指标[58]以及与来自突发组装队列的标准统计和状态信息组合使用，IP-PAC可做出阈值调整。如果到达流量是高度突发的，即有一个大的Hurst参数值，则汇聚过程的时间将减少（如果缓冲几乎满的话）或增加（如果缓冲几乎空的话)。以一种类似的方式，如果流量不是突发的，即给出一个低的Hurst参数值，则汇聚时间将被设置为它的默认值。增加/减少因子是上次估计的Hurst参数值的一个函数。假定也存在网络反馈作为链路上的预测网络负载，则汇聚分组的尺寸可被设置为满足网络状态的当前条件。分组的尺寸是在网络约束的基础上设置的，这些约束如链路MTU和路径MTU[59]，见IPv6定义中的规定[55]。如果可能，一条分组的最大尺寸可达到一个巨型数据报的尺寸[38]，高达4GB。可使用ICMPv6[60]组织网络反馈的格式和规程，以及IP-PAC状态通信。

图10.11给出了由一台IP-PAC机器产生的一条样例分组。注意，初始IP首部(显示为黄色或灰色阴影）具有有关分组总长度的信息。IP-PAC知道第一个首部后跟的数据，也是一个IP首部，并由此解释接下来的分组，将之从净荷中去除，并转发到它的目的地。直到净荷为空之前，都要执行这个过程。

解决从净荷中抽取分组的问题，有其他方式。一种可能是以分组首部的解释来映射汇聚过的分组。因为仅有第一个首部的位置是已知的，所以机器可在第一个首部带有的信息基础上计算后跟的首部，并在所有分组被抽取之前，按顺序抽取分组。

注意，每个目的地存在一个队列，是一种简化，更准确地说，为处理与不同类型流量或QoS有关的问题，IP-PAC的每个目的地可能有几个队列。

10.7　全IP光突发交换网络

一个WDM上传输IP的网络通常是这样实现的，它考虑到IP分组将会被成帧，做法是将分组汇聚到一个SONET/SDH帧或一个以太网帧内。一些架构进一步提出，使用多协议标记交换（MPLS）和多协议λ交换（MPλS)，如标记光突发交换(LOBS）架构的情形[61]。在这种情形中，MPLS消息被用来管理OBS网络，暗示这样的架构将有利于真正的光分组交换的未来实现。

全IP OBS网络是这样一个概念，它依赖于在一个OBS网络上直接传输IP，这

意味着构成分支流量的各IP分组将在OBS网络内被汇聚和交换。

通过回顾WDM上传输IP的概念，首先介绍OBS网络的概念。在WDM上传输IP的一个网络上执行数据的传输，有两种基本方式，一种方式使用可重新配置的WDM节点，如WDM网络上叠加SONET再叠加IP做法中的方式，另一种方式使用交换式WDM工作节点。工作在交换式方式中的一个节点，简单地是支持其OXC矩阵快速重新配置的一个节点。表10.1给出了WDM网络工作的这两种类型的比较。

表10.1 WDM技术比较（摘自参考文献［15］）

	可重新配置的WDM	交换式WDM		
交换技术	电路	突发	标记	分组
控制消息	带外	带外	带外/带内	带内
流量粒度	大型	中等	中等	小型
数据交换单元	波长或光纤	数据突发	流	分组
数据传输	电路	应需电路	虚信道	下一跳

一个OBS网络工作过程如下，当突发要从入口点穿过网络到出口点所需的时间期间，创建一条临时电路。在入口节点组装突发，并在出口点拆装。当路径中的每个节点解释包含在一条控制分组［在分组出现在一条带外数据信道（λ）之前发送的］中的信息时，完成电路的创建。图10.14给出了组成一个OBS网络的单元。从中看到3个核心节点（由其相应的OXC和SE表示）和链路（光纤）（连接各节点），给出数据信道（粗线）和控制信道（细线）。

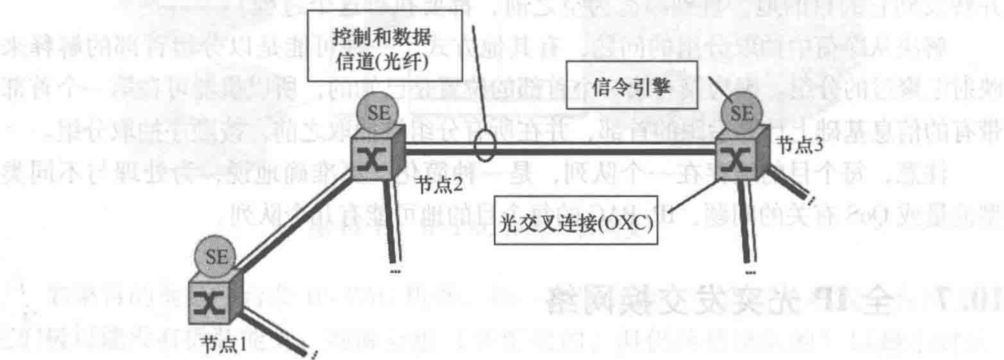


图10.14 一个OBS架构（部分）的方案（摘自参考文献［1］）

OBS网络集成具有特定和差异性功能的两种类型的节点，即边缘节点和核心节点[62,63]。边缘节点没有OXC，由此不实施突发交换，但它们负责分支数据单元（分组、信元、帧等）汇聚成突发（突发组装）和从突发中检索（或解聚）出数据单元（拆装）。边缘节点与OBS核心节点（通过一条光链路）接口，并与客户网络接口。

核心节点以完全格式无关的方式实施突发交换，原因在于在节点中的资源以突发通过的必要时间进行配置，即核心节点是不知道数据突发的中转的。

图10.15给出了一个样例配置，包括连接到一个OBS传输网络的两个客户网络。

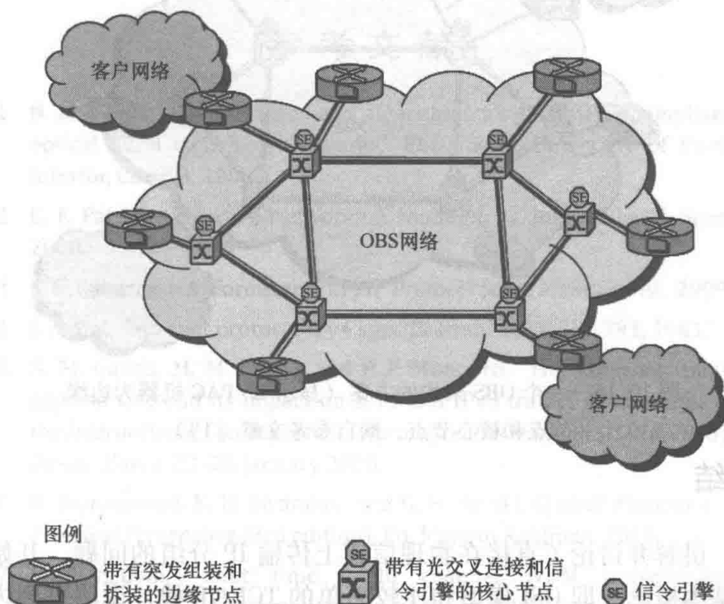


图10.15 一个OBS架构的方案（显示边缘节点和核心节点，摘自参考文献［1］）

这里可看到，客户网络与一个边缘节点接口（虽然它们可与几个边缘节点接口，指在这种情形中，客户网络可有用于传输网络的两台网关），边缘节点与一个核心节点接口。同样这里可能存在这样一种配置，其中一个边缘节点与一个以上的核心节点接口，这取决于在边缘节点处存在的WDM接口数。因为边缘节点必须实现突发组装，所以它必须为OBS云上的每个可能目的地（对于每个可能的出口节点）和每个被处理的流量类创建一个突发组装队列。在图10.15所示的例子中，有7个边缘节点，意味着有7个客户网络。如果每个节点处理3类流量服务［如优惠、尽力而为和慢速］，那么在每个入口边缘节点处将有6个可能的目的地，对每个目的地，有3个队列，所以每个入口节点将需要维护18个突发组装队列。另外，因为对每个接收的突发，边缘节点需要一个突发拆装队列，所以突发拆装队列的创建需要是动态的。

边缘节点作为突发组装和拆装单元，在OBS中由许多流量源和格式形成一次突发，这取决于哪种类型的客户网络及其哪些接口连接到边缘节点。

在分支网络为IP网络的一个场景中，可以IP-PAC机器替换边缘节点，由此将如图10.15所示的配置转换为如图10.16所示的配置。对于这种配置，客户网络是IP（v4、v6、vFuture）客户网络，IP-PAC机器实施突发组装和IP分组拆装为相应格式的突发。在这样一个场景中使用IP-PAC机器，配置成一个全IP OBS网络。

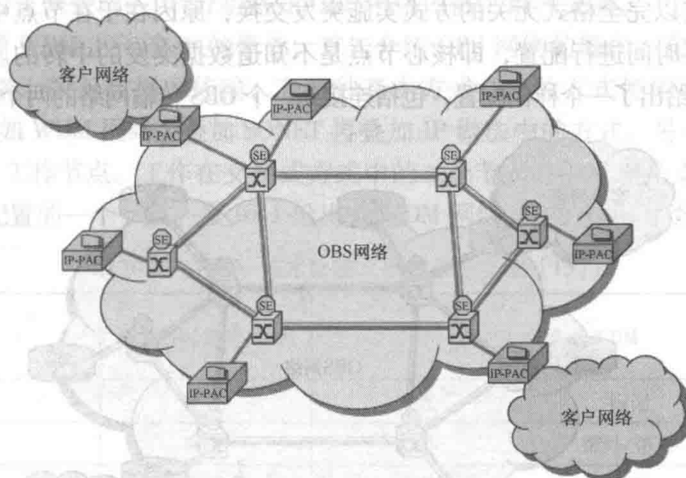


图 10.16 一个 OBS 架构的方案（显示 IP-PAC 机器为边缘节点和核心节点，摘自参考文献［1］）

10.8 小结

在本章，讲解并讨论了直接在物理信道上传输 IP 分组的问题。开始时，介绍现代网络的基本概念，即 OSI 模型和比较简单的 TCP/IP 栈，说明直到内容实际上被放置到传输媒介［如一条光纤，或以太（用于无线传输）］上之前，在一台计算机上由一个人产生的内容要如何经过变换。将这些连续的变换识别为封装过程，论证每个附加的信封（首部）都服务于一个目的，并帮助用户的内容被成功地交付到目的主机和目的应用。

也将例子扩展，包括长距离数据传输，描述了 WDM 技术上传输 IP 的最新技术现状。详细描述了这种技术方便使用的基础依据，并扩展 WDM 网络上传输 IP 的主题，包括基本电路交换技术和当前成帧方法，讲解了 SONET 和 SDH。也专门描述了控制和管理方式方面的一项功能，这项功能当前实现在这样的 SONET 或 SDH 网络之中。

提出 OBS 网络，作为 WDM 网络上传输 IP 的一个特例，介绍时辅以论据，即支持非常短寿命电路的创建，这样这些网络可克服在 WDM 架构上电路交换式光路径 IP 的低效问题。

也以与光网络无关的一种方式，讲解了 IP 分组汇聚的概念，即从一个媒介无关的观点讨论了这个概念。讨论了这样一种机器的工作模式及其约束。

最后，将 OBS 和 IP-PAC 的概念放在一起，设计一种全光 OBS 网络。在这个场景中，边缘节点是 IP-PAC 机器，核心节点是常规的 OBS 节点或具有特殊功能特征的节点，这项特征支持更高级 OBS 架构的实现。

在本章中，假定当在传输或核心网中传输分组时不必要的封装和成帧，虽然作为由汇聚带来的统计复用效用的结果，这可受益于交换操作的减少，但也增加了网

络工作，原因是需要拆装 SDH 帧，以便检索各组成元素。

受益于 OBS 网络的速度和可靠性，全 IP OBS 网络的建议结果是可得到由泛在 IP 客户网络传输数据的一种更快速和更便利的方式。

参考文献

1. N. M. Garcia, "Architectures and algorithms for IPv4/IPv6-compliant optical burst switching networks," PhD thesis, University of Beira Interior, Covilhã, 2008.
2. C. F. Patrick Ciccarelli, *Networking Foundations*, John Wiley & Sons, 2008.
3. S. C. F. Behrouz A. Forouzan, *TCP/IP Protocol Suite*, McGraw-Hill, 2009.
4. J. Postel, "Internet protocol IPv4 specification," IETF RFC 791, 1981.
5. N. M. Garcia, M. M. Freire, and P. P. Monteiro, "The Ethernet frame payload size and its impact on IPv4 and IPv6 traffic," *Proceedings of the International Conference on Information Networking* (*ICOIN 2008*), Busan, Korea, 23–25 January 2008.
6. R. Ramaswami, K. N. Sivarajan, and G. H. Sasaki, *Optical Networks, A Practical Perspective* (3rd edition), Ed. Morgan Kaufman, 2010.
7. PhysOrg.Com, "First time 1,000 channel WDM transmission demonstration in an installed optical fiber," 2005, accessed July 13, 2007, http://www.physorg.com/ news3316.html.
8. C. Qiao and M. Yoo, "Optical burst switching (OBS)—a new paradigm for an optical Internet," *Journal of High Speed Networks*, **8(1)**, 69–84, January 1999.
9. Y. Chen, C. Qiao, and X. Yu, "Optical burst switching: a new area in optical networking research," *IEEE Network*, **18(3)**, 16–23, May–June 2004.
10. T. Battestilli and H. Perros, "Optical burst switching: a survey," North Carolina State University, *NCSU Computer Science Technical Report*, TR-2002-10, July 2002.
11. V. Puttasubbappa, "Optical burst switching: challenges, solutions and performance evaluation," PhD thesis, North Carolina State University, Raleigh, 2006.
12. A. Gumaste and T. Antony, *Optical Network Design and Implementation: Introduction to First Mile Access Technologies*, Cisco Press, 2004.
13. I. Chlamtac, A. Ganz, and G. Karmi, "Lightpath communications: an approach to high bandwidth optican WAN's," *IEEE Transactions on Communications*, **40(7)**, 1171–1182, July 1992.
14. Cisco Systems Inc., "Cisco XR 12000 and 12000 series [Cisco 12000 series routers]," Cisco Systems, 2006, accessed July 20, 2011, http://www.cisco.com/en/US/products/hw/routers/ps167/products_qanda_item0900aecd8027c915.shtml.

15. K. H. Liu, *IP Over WDM*, John Wiley & Sons, West Sussex, 2002.
16. Fiberdyne Labs, "Dense wave division multiplexing (DWDM) ITU grid C-band, 100 GHz spacing," 2011, accessed July 20, 2011, http://www.fiberdyne.com/products/itu-grid.html.
17. B. Rajagopalan, J. Luciani, and D. Awduche, "IP over optical networks: a framework," IETF RFC 3717, 1998.
18. A. Fredette and J. Lang, "Link management protocol (LMP) for dense wavelength division multiplexing (DWDM) optical line systems," IETF RFC 4209, 2005.
19. J. Case, M. Fedor, M. Schoffstall, and J. Davin, "A simple network management protocol," IETF RFC 1098, 1990.
20. K. McCloghrie and M. Rose, "Management information base for network management of TCP/IP-based internets: MIB-II," IETF RFC 1213, 1991.
21. Object Management Group, "Catalog of OMG CORBA (R)/IIOP (R) specifications," accessed July 13, 2011, http://www.omg.org/technology/documents/corba_spec_ catalog.htm, 2011.
22. E. F. Haselton, "A PCM frame switching concept leading to burst switching network architecture," *IEEE Communications Magazine*, **21(6)**, 13–19, September 1983.
23. S. R. Amstutz, "Burst switching—an introduction," *IEEE Communications Magazine*, **21(8)**, 36–42, November 1983.
24. S. R. Amstutz, "Burst switching—an Update," *IEEE Communications Magazine*, **27(9)**, 50–57, September 1989.
25. J. S. Turner, "Terabit burst switching," *Journal of High Speed Networks*, **8(1)**, 3–16, January 1999.
26. J. Y. Wei and R. I. McFarland, "Just-in-time signaling for WDM optical burst switching networks," *Journal of Lightwave Technology*, **18(12)**, 2019–2037, December 2000.
27. I. Baldine, G. Rouskas, H. Perros, and D. Stevenson, "JumpStart—a just-in-time signaling architecture for WDM burst-switched networks," *IEEE Communications Magazine*, **40(2)**, 82–89, February 2002.
28. J. Teng and G. N. Rouskas, "A comparison of the JIT, JET, and horizon wavelength reservation schemes on a single OBS node," *Proceedings of WOBS 2003*, Dallas, Texas.
29. P. Schlüter, "Aggregation of IP flows," Siemens AG, 2000, accessed January 15, 2006, http://mr-ip.icm.siemens.de/mr/traffic/TR/flow-agg.pdf.
30. S. Malik and U. Killat, "Impact of burst aggregation time on performance in optical burst switching networks," *Proceedings of Optical Fibre Communications Conference (OFC 2004)*, 2, 2, Los Angeles, CA, February 23–27, 2004.

31. T. Ferrari, "End-to-end performance analysis with traffic aggregation," *Computer Networks*, **34(6)**, 905-914, 2000.

32. K. Dolzer, "Assured horizon—an efficient framework for service differentiation in optical burst switched networks," *Proceedings of SPIE Optical Networking and Communications Conference (OptiComm 2002)*, Boston, MA, July 30-31, 2002.

33. A. Zapata and P. Bayvel, "Impact of burst aggregation schemes on delay in optical burst switched networks," *Proceedings of IEEE/LEOS 2003*, Tucson, AZ, October 26-30, 2003.

34. A. Sridharan, S. Bhattacharyya, C. Dyot, R. Guérin, J. Jetcheva, and N. Taft, "On the impact of aggregation on the performance of traffic aware routing," *Proceedings of International Teletraffic Congress*, Salvador da Bahia, Brazil, December 2001.

35. X. Mountrouidou and H. Perros, "Characterization of burst aggregation process in optical burst switching," *Proceedings of Networking 2006*, **1**, 752-764, Coimbra, Portugal, May 15-19, 2006.

36. N. M. Garcia, P. P. Monteiro, and M. M. Freire, "Burst assembly with real IPv4 data—performance assessment of three assembly algorithms," *Next Generation Teletraffic and Wired/Wireless Advanced Networking 2006, Lecture Notes in Computer Science, LCNS 4003*, Ed. St. Petersburg, Russia; Springer-Verlag, Berlin, Heidelberg, pp. 223-234, May 2006.

37 N. M. Garcia, P. P. Monteiro, and M. M. Freire, "Assessment of burst assembly algorithms using real IPv4 data traces," *Proceedings of the 2nd International Conference on Distributed Frameworks for Multimedia Applications (DFMA'2006)*, 173-178, Pulau Pinang, Malaysia, May 14-17, 2006.

38. D. Borman, S. Deering, and R. Hinden, "IPv6 jumbograms," IETF RFC 2675, 1999.

39. V. M. Vokkarane, K. Haridoss, and J. P. Jue, "Threshold-based burst assembly policies for QoS support on optical burst-switched networks," *Proceedings of SPIE Optical Networking and Communications Conference (OPTICOMM 2002)*, 125-136, Boston, MA, July 29-August 1, 2002.

40. A. Ge and F. Callegati, "On optical burst switching and self-similar traffic," *IEEE Communications Letters*, **4(3)**, 98-100, March 2000.

41. X. Yu, Y. Chen, and C. Qiao, "Performance evaluation of optical burst switching with assembled burst traffic input," *Proceedings of IEEE Global Telecommunications Conference (GLOBECOM 2002)*, **3**, 2318-2322, Taipei, November 11-17, 2002.

42. M. C. Yuang, J. Shil, and P. L. Tien, "QoS burstification for optical burst switched WDM networks," *Proceedings of Optical Fiber Communication Conference (OFC 2002)*, Anaheim, CA, March 17-22, 2002.

43. J. Postel, "Assigned numbers," Internet Engineering Task Force, 1981.
44. H. Ning, "IPv6 test-bed networks and R&D in China," *Proceedings of the International Symposium on Applications and the Internet Workshops (SAINTW'04)*, Tokyo, Japan, January 26–30, 2004.
45. T.-Y. Wu, H.-C. Chao, T.-G. Tsuei, and E. Lee, "A measurement study of network efficiency for TWAREN IPv6 backbone," *International Journal of Network Management*, **15(6)**, 411–419, November 2005.
46. D. G. Waddington and F. Chang, "Realizing the transition to IPv6," *IEEE Communications Magazine*, **40(6)**, 139–144, June 2002.
47. J.-M. Uzé, "Enabling IPv6 services in ISP networks," Presented to International Workshop on IPv6 Testing Certification and Market Acceptance, Brussels, Belgium, September 22, 2003.
48. C. E. Hopps, "Routing IPv6 with IS-IS," IETF draft, 2003.
49. J. Harrison, J. Berger, and M. Bartlett, "IPv6 traffic engineering in IS-IS," IETF draft, 2005.
50. K. Cho, M. Luckie, and B. Hufftaker, "Identifying IPv6 network problems in the dualstack world," *Proceedings of SIGCOMM'04 Workshops*, Portland, Oregon, September 3, 2004.
51. Y. Adam, B. Fillinger, I. Astic, A. Lahmadi, and P. Brigant, "Deployment and test of IPv6 services in the VTHD network," *IEEE Communications Magazine*, 0163(6804-08), 98–104, January 2004.
52. J. Bound, "IPv6 deployment next steps & focus (infrastructure!!!)," Presented to IPv6 US Summit, Arlington, VA, December 8–11, 2003.
53. P. Hovel, "Barriers to IPv6 deployment," *European Commission IPv6 Task Force Report*, June 6, 2003.
54. Cisco Systems, "IPv6 deployment strategies," 2002, accessed September15,2007,http://www.cisco.com/univercd/cc/td/doc/cisintwk/intsolns/ipv6_sol/ ipv6dswp.pdf.
55. S. Deering and R. Hinden, "Internet protocol, version 6 (IPv6) specification," IETF RFC 2460, 1998.
56. M. M. Freire, N. M. Garcia, M. Hajduczenia, P. P. Monteiro, and H. Silva, "Method for aggregating a plurality of data packets with different IP formats and machine for performing said method "Patent filed in WIPO, Ref. International Patent WO 2007/118594-A1, 2007.
57. N. Nagatsu, S. Okamoto, and K.-I. Sato, "Optical path cross/connect system scale evaluation using path accommodation design for restricted wavelength multiplexing," *IEEE Journal on Selected Areas in Communications*, **14(5)**, 893–901, June 1996.
58. M. Hajduczenia, N. M. Garcia, P. Monteiro, H. Silva, and M. M. Freire, "Monitoring method and apparatus of processing of a data stream with high rate/flow "patent filed in European Patent Filed in the European Patent Office, Ref. Patent Number 05023580.3-2416, April 20, 2005.

59. J. McCann, S. Deering, and J. Mogul, "Path MTU discovery for IP version 6," IETF RFC 1981, 1996.

60. A. Conta, S. Deering, and M. Gupta, "Internet control message protocol (ICMPv6) for the Internet protocol version 6 (IPv6) specification," IETF draft on RFC 2463, 2004.

61. C. Qiao and J. Staley, "Method to route and re-route data in OBS/LOBS and other burst switched networks," Patent filed in United States Patent Office, Ref. 6, November 2003.

62. C. Kan, H. Balt, S. Michel, and D. Verchère, "Network-element view information model for an optical burst core switch," *Proceedings of the International Society for Optical Engineering—SPIE Asia-Pacific Optical Wireless Communications* (*APOC 2001*), **4584**, 115–125, Beijing, China, November 12–16, 2001.

63. C. Kan, H. Balt, S. Michel, and D. Verchère, "Information model of an optical burst edge switch," *Proceedings of the IEEE International Conference on Communications* (*ICC 2002*), **5**, 2717–2721, New York, April 28–May 2, 2002.

第 11 章　WiMAX 上的 IPv6

Jayateertha G. M., B. Ashwini

11.1　引言

在过去 10 年，无线通信和数据通信（因特网）在技术演进和发展方面经历了显著增长。在针对各种多媒体应用的厂商和服务提供商的新市场经营和商务机遇方面，这两个世界都有巨大的增长前景，使它们趋向融合。所以，在无线通信上提供高速因特网连接，是日渐成熟的融合技术的主要可圈可点的任务。出现了不同的技术，提出通过无线通信连接到因特网，如 IEEE 802.11（Wi-Fi）、第三代（3G）技术和全球微波接入互操作性（WiMAX）（IEEE 802.16）。在这些技术中，WiMAX 是宽带无线通信的事实标准。WiMAX 支持固定/移动用户无线宽带服务的泛在交付。当前移动 WiMAX 技术主要基于 802.16e 标准，该标准规范了一个正交频分多址（OFDMA）空中接口，并提供移动性支持。采用灵活的带宽分配、多种内建的服务质量（QoS）类型支持和覆盖距离为 50km、标称数据速率高达 100 Mbit/s，WiMAX 为多媒体服务部署准备了条件，这些服务如因特网协议上的话音（VoIP）、视频点播（VoD）、视频会议、多媒体聊天和移动娱乐。另外，具有高速连接的因特网的巨大增长，这是由于采用传输控制协议/因特网协议（TCP/IP）的因特网协议版本 4（IPv4）的快速发展，以及高速路由器和内建的 QoS 机制，使 IP 成为传输和路由多媒体实时分组的事实标准。所以，从运营商的角度看，特别是支持多媒体接入类型（如 3G 和 WiMAX）的那些运营商，利用 IP 作为传输和路由分组的方法，就是有道理的。这些新的无线技术使用 IP 来维护连通性的事实，可能导致 IP 地址稀缺的问题。考虑到无线工业界的指数增长，接下来的最近时间引入各种无线设备，IPv4 地址空间就不足以将它们都连接起来。为解决这个问题，设计了因特网协议版本 6（IPv6）替代 IPv4，它将 IP 地址长度从 32 比特扩展到 128 比特，由此提供了一个巨大的地址空间。在提供一个巨大地址空间的同时，IPv6 也被设计为处理因特网的增长率，并以其内建的 QoS 和安全特征来处理对服务、移动性和端到端安全的紧迫需求。IPv6 最有趣的新特征是无状态自动配置机制。利用这种机制，一台启动设备通过使用它的媒介访问控制（MAC）地址或一个私有的随机数，自动配置它自己的全局唯一 IP 地址。这种机制依赖于低层容量来处理组播通信。另外，WiMAX 基于一种点到多点架构，其中在两个静态/移动站之间的 MAC 层处没有授权直接通信，而是所有通信都开始并终止于基站（BS）。由此，

在 WiMAX 中不支持 MAC 层的组播通信，这就为在 WiMAX 架构上部署 IPv6 产生了某种推动力。WiMAX 论坛［联网组（NWG）］，着手处理这项挑战，攻克这个部署问题，它为在 WiMAX 上部署 IPv6（还有一些问题）建议了一个模型。事实上这是本章的主题。本章后面内容安排如下：11.2 节给出 IEEE 802.16e 标准概述；11.3 节描述 WiMAX 网络架构；11.4 节讨论 IPv6 自动配置概念；11.5 节基于 WiMAX 论坛提出的模型和一些解决方案，讨论在 WiMAX 上部署 IPv6；11.6 节讨论在 WiMAX 上部署 IPv6 预见到的一些问题。

11.2 WiMAX 技术概述

本节意图提供当前移动 WiMAX 技术的一个高层概述，重点是与目前讨论主题有关的空中接口［物理层（PHY）和 MAC 层］。在下一节将讨论相关的 WiMAX 技术网络架构。

WiMAX 技术的空中接口是基于 IEEE 802.16 标准的。这些标准为 PHY 和 MAC 层功能提供规范。特别地，当前的移动 WiMAX 技术是基于 IEEE 802.16e 的，该标准规范了一个 OFDMA 空中接口，并提供对移动性的支持。移动 WiMAX 技术的网络规范目标是针对大范围的 IP 服务做过优化的一个端到端全 IP 架构，这些服务包括网络互操作性。WiMAX 论坛的 NWG 负责这些规范，这些规范补充 IEEE 802.16e 规范。图 11.1 给出了移动 WiMAX 技术的组成，常被称作发行版 1.0 概要[3]。

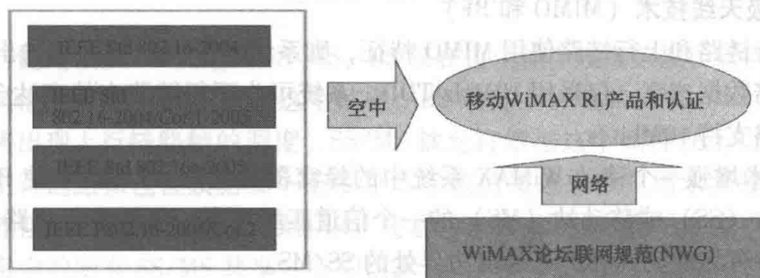


图 11.1　移动 WiMAX 发行版本 1.0

空中接口规范由 4 个相关的 IEEE 802.16 宽带无线接入标准组成。WiMAX 系统概要由 5 个子概要组成，即 PHY、MAC、无线电、复用模式和功率类。

11.2.1 物理层

IEEE 802.16e 的 PHY[15]工作在 2.5 GHz 频带，具有像 OFDMA 和多输入多输出（MIMO）的高级特征，并以非视距（N-LOS）通信支持移动性。PHY 支持 30～130Mbit/s 范围间的数据速率，这取决于工作带宽以及调制和编码方案。20MHz、

25MHz或28MHz的信道带宽与正交移相键控（QPSK）、16正交幅度调制（QAM）和64-QAM作为调制技术一起使用，具体使用哪种调制技术取决于信道条件。系统使用0.5ms、1ms、2ms的帧尺寸进行传输，针对下行链路和上行链路传输，一个帧被分成子帧。下一节简短地讨论了WiMAX技术的一些关键PHY特征。

1. OFDM作为一项接入技术

OFDMA是基于多址方案的正交频分复用（OFDM）。OFDMA以相对简单的接收转发器结构，不仅在N-LOS多信道中展示了卓越的性能，而且以合理的复杂度支持高级天线技术（MIMO）的可行实现。它支持依据时间和子信道化的频谱资源的高效使用。通过灵活地调整快速傅里叶变换（FFT）尺寸、具有一个固定符号时长的信道带宽和子载波间隙，在这方面来说，OFDMA是可扩展的，它可解决各种频谱需要。

2. 时分复用

虽然基础IEEE 802.16标准包含TDD和频分复用（FDD），但移动WiMAX发行版本1.0概要[3]仅规范时分复用（TDD）作为复用模式。事实上，相比FDD，TDD是较好定位于移动因特网服务的。首先，因为因特网流量的下行链路流量超过上行链路流量，所以因特网流量是非对称的。由此，采用常规FDD，相同的下行链路和上行链路容量不能提供最优的无线电资源使用率。采用TDD，可依据他们的服务需要，调整上行链路和下行链路比率。此外，相比FDD，TDD更适用于高级天线技术，如自适应天线系统（AAS）或波束形成（BF），这是由于上行链路和下行链路之间的信道反比关系导致的。

3. 高级天线技术（MIMO和BF）

为下行链路和上行链路使用MIMO特征，则系统可支持单输入单输出（SISO）速率的双倍数据速率。仅采用10MHz TDD，系统可为下行链路支持高达37Mbit/s，为上行链路支持10Mbit/s。

BF技术增强一个移动WiMAX系统中的蜂窝覆盖范围。BF机制使BS能够形成到用户站（SS）或移动站（MS）的一个信道匹配波束，从而上行链路和下行链路信号能够可靠地到达和离开蜂窝边界处的SS/MS。

4. 全移动性支持

采用诸如混合自动重发请求（HARQ）的高级特征，移动WiMAX可支持以高速路速度运动的车辆。HARQ技术有助于缓解快速衰落信道效应和干扰振荡，由此以组合的增益和时间多样性改善整体性能。

5. 灵活的频率重用

具有蜂窝特定子信道化、低率编码以及功率放大和去放大特征的移动WiMAX，可在高干扰受限的条件下，改善整体系统的性能。它也支持灵活频率重用的实时应用。一般而言，频率重用被应用到靠近蜂窝中心的SS/MS，而一部分频率被用在蜂窝边缘的SS/MS。这有助于降低严重的共信道干扰。

11.2.2 MAC 层

移动 WiMAX（IEEE 802.16e）的 MAC 层技术包括为高效率和灵活性而提供的多项功能特征。在 MAC 层，移动 WiMAX 网络被看作一条点到多点连接，这是在 WiMAX 上部署 IPv6 的主要问题。就这方面而言，传输和控制的所有数据通信都处在单向连接之中。在接下来的一节将简短地讨论移动 WiMAX 技术的一些关键 MAC 功能特征。

1. 基于调度连接的数据传输

移动 WiMAX 技术为面向连接的服务提供一个环境。无论传输还是控制的数据通信，都在单向连接中传输。出于这个目的，将无线媒介分成上行链路帧和下行链路帧。BS 应需将资源授予 SS/MS，并在上行链路上调度流量。MAC 层支持 TDD 和 FDD，其中 TDD 在时间上隔离上行链路和下行链路，而 FDD 在频率上将它们隔离开。图 11.2 给出了通用的 TDD 超帧结构。依据不同的物理概要，帧尺寸也是变化的，它在上行链路和下行链路之间的分隔，可依据流量分布加以调整。

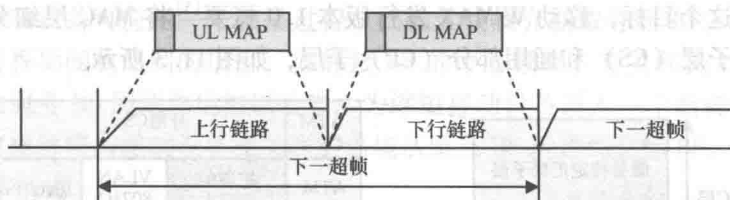


图 11.2　TDD 超帧结构

下行链路帧从 BS 发送到 SS/MS，BS 仅在这个帧上发送数据，所以这些时间槽是不与其他 SS/MS 共享的。在这个帧中，BS 通过上行链路映射（UL MAP）消息发送即将出现上行链路帧的调度。SS/MS 被允许依据这个调度在上行链路上进行传输。上行链路帧包含数据或一条带 QoS 参数的每连接带宽请求消息，这些连接以一个连接标识符（CID）加以标识。注意，来自 BS 的数据是在空中广播的，由那个蜂窝中的所有 SS/MS 共享（空中信道），仅有与 CID（包括在传输的帧中）关联的 SS/MS 才可访问内容。由此可注意到，在两个 SS/MS 之间没有直接的通信。如果一个 SS/MS 希望与其他 SS/MS 通信，则必须通过 BS 进行。

2. 灵活的带宽分配机制

带宽分配机制是基于实时带宽请求的，这些请求依据连接由 SS/MS 传输到 BS。这些请求可使用一种基于碰撞的机制在上行链路帧上发送，或与数据消息一起捎带。BS 基于这些请求和连接的 QoS 参数，实施资源分配。

3. 依据连接的服务分类和质量

依据应用的 QoS 服务需求和流量模式，移动 WiMAX 发行版本 1.0 概要[3]在 MAC 层对应用分类。QoS 服务需求是（诸如）带有严格时延和带宽需求的实时多

媒体应用或带有最小保障带宽的尽力而为（BE）应用。流量模式可以类似固定/可变数据分组、周期性的/非周期性的等。至此，最初的 IEEE 802.16 标准[14]定义了 4 个流量类，IEEE 802.16e 添加了第五个流量类。这些流量类如下：

1）非请求授权服务（UGS）：支持恒定比特率（CBR）服务，如 T1/E1 模拟和 VoIP（在没有静默抑制条件下）。

2）实时查询服务（rtPS）：在周期性的基础上，支持带有可变尺寸数据的实时服务，如 MPEG 和带有静默抑制的 VoIP。

3）扩展的实时查询服务（ertPS）：最近由 IEEE 802.16e 标准引入的。它组合使用 UGS 和 rtPS，即确保周期性的非请求授权，但授权尺寸可由请求加以改变。

4）非实时查询服务（nrtPS）：支持非实时服务，这些服务在常规基础上要求可变尺寸的数据突发，如文件传输协议（FTP）服务。

5）BE：用于不要求 QoS 的应用，如超文本传输协议（HTTP）。

要求一条连接的每个 SS/MS，都需要包括与上述各类相关的 QoS 参数。

4. 对不同网络服务的支持

为达到这个目标，移动 WiMAX 发行版本 1.0 概要[3]将 MAC 层细分为两个子层，即汇聚子层（CS）和通用部分（CP）子层，如图 11.3 所示。

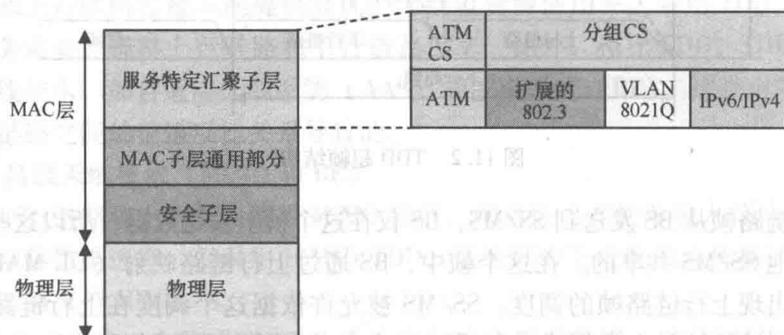


图 11.3 移动 WiMAX MAC 层

CS 将各种传输层流量映射到核心 MAC 通用部分子层。换句话说，这个层处理（例如）异步传递模式（ATM）信元、IP 分组、以太网分组和虚拟局域网（VLAN）分组的汇聚，以便 MAC 可支持 ATM 服务和分组服务。CS 依据到达信元/分组的流量类型（如话音、网络冲浪、ATM、CBR），对到达信元/分组（称作 SDU）分类，并使用一个 32 比特服务流 ID（SFID），将信元/分组指派到服务流。在如今 WiMAX 产品中的最重要的 CS 有因特网协议汇聚子层（IP CS）和以太网 CS。

对于一个 IP CS，可依据 IP 地址（源和目的地）、传输端口［TCP 或用户数据报协议（UDP）］和服务类型字段对各 SDU 进行分类。对于一个以太网 CS，在以太网地址（源和目的地）和用户记录的优先级基础上，对各 SDU 分类。

通用部分子层独立于传输机制，并主要负责将各 SDU 分片和分段为 MAC 协议数据单元（PDU）、句柄、QoS 控制以及 MAC PDU 的调度和重传。

一旦接纳服务流，那么它就被映射到一个唯一的 16 比特 MAP 连接标识符（CID），CID 处理服务流的 QoS 需求。服务流刻画为它的 QoS 参数，这些参数描述它的延迟、抖动、吞吐量需求等。在指派一个 CID 之后，服务流被转发到一个合适的队列，分组调度器从队列中检索分组，并将其传输到网络。使用自适应突发概要法，每项服务被指派到一个合适的 PHY 配置，以便处理该服务。

5. MAC 开销降低

移动 WiMAX 技术支持通用首部抑制（PHS）和 IP 首部压缩，即因特网工程任务组（IETF）标准 ROHC。PHS 可用于任何格式的分组，如以太网上的 IPv4 或 IPv6。如果流量的相当部分具有相同的首部，如 IP 或以太网，则 PHS 是比较有效的。为降低首部开销，PHS 机制以一个短的语境标识符替换首部的重复部分。

6. 移动性支持：切换

就切换延迟而言，移动 WiMAX 技术提供了高度优化过的技术。当远离服务 BS 时，一台移动设备可应用一个扫描过程，扫描无线媒介获取邻居 BS。在切换中可使用扫描过程期间收集到的信息，如邻接 BS 的中心频率。有关邻接 BS 的这种信息，也可由服务 BS 周期性地加以公告。为缩短移动设备进入一个新蜂窝所需的时间，网络也能够将与移动设备相关联的语境从服务 BS 传递到目标 BS。

7. 功率节省

SS/MS 的睡眠模式和空闲模式规程可用于功率节省。在睡眠模式，移动设备保持与 BS 的注册状态，但可关闭它的一些电路，以便降低功率消耗。当移动设备长时间没有流量时，它就进入空闲模式。在这种模式中，移动设备没有注册到任何 BS。为恢复移动设备和网络之间的流量，网络使用一个寻呼规程。

8. 安全

MAC 层的安全子层在移动设备和网络之间提供基于扩展认证协议（EAP）的双向认证。为预防对所传递数据的非授权访问，则使用高效的加密规程。通过将一个基于数字证书的 SS/MS 设备认证添加到加密密钥管理协议，基本安全机制得以改进。

9. 对下行链路组播和广播服务的支持

即使当一台 MS 处于空闲模式时，WiMAX 技术的组播和广播服务（MBS）也支持它接收组播和广播数据。这项功能特征的最流行例子是 TV 广播到移动终端的例子。

11.3 WiMAX 网络架构

移动 WiMAX 网络规范[2]目标为针对大范围 IP 服务做过优化的一种端到端全 IP 架构。2.5G 和 3G 系统采用物理系统设计原则，与此不同，WiMAX 论坛的

NWG，基于遵循因特网演进路线的功能设计原则定义了一个规范，其互操作能力是基于没有PHY的协议和规程的。由此，NWG定义了由功能实体组成的一个功能性网络架构，大量使用IP和IETF标准协议。

主要焦点是支持移动设备的IP接入。在网络中的所有交互实体之间，假定具有IP连通性。客户端设备的联网功能由标准IP协议组成，如动态主机配置协议（DHCP）、移动IP、EAP等。当一台移动设备从一个接入服务网络（ASN）移动到另一个ASN，跨越IP子网边界时，使用移动IP作为数据重定向的机制。在网络侧，通过诸如DHCP、认证/授权/计费（AAA）等的标准IP，提供IP地址池管理。在因特网演进的线路中，它遵循跨联网实体分解协议的策略，这些实体支持互操作能力，同时为厂商和运营商提供灵活的实现选择。

端到端移动WiMAX技术采用如下设计原则：

1）针对IP服务优化的无线电ASN：在11.2节中讨论的所有功能特征被看作无线电基础，它同时交付大范围的IP服务，如实时话音、单播、组播多媒体和基于TCP的服务。

2）IP互联的ASN：采用功能自治，支持更扁平的ASN架构；基于IP的互连接能力，支持固有的冗余和扩展性；基于因特网设计原则，功能架构支持不同的物理实现和拓扑。

3）ASN与连通及应用服务网络（CSN）的逻辑分离：对在两个或多个连接服务提供商间共享ASN的支持；也支持这样的一个连接服3）务提供商，它在由两个或多个运营商所部署的ASN上提供宽带服务。

4）由ASN组成的网络：对无线接入部件和核心IP服务功能之间基于IP的开放接口的支持，这样就支持到未来移动宽带接入技术的独立演进和过渡。

网络参考模型

在高层的移动WiMAX网络架构[2]被表示为一个网络参考模型（NRM），它定义了关键的功能实体和参考点（RP），在其上可构建互操作网络框架。一个NRM由逻辑实体组成，即MS、ASN、CSN及其通过RP R1～R5的交互（见图11.4）。在高层，NRM将WiMAX网络架构在两个商务实体之间做了划分，即网络接入提供商（NAP）和网络服务提供商（NSP）。一个NAP以一个或多个ASN提供WiMAX无线电接入基础设施，而一个NSP为WiMAX用户（控制CSN）提供IP连接和WiMAX服务。

1. 网络功能实体

MS、ASN和CSN，每个都代表了功能的一个逻辑分组，如下描述。

（1）移动站（MS）/用户站

它是一个广义的移动设备集合，在一台或多台主机和WiMAX网络之间提供无线连接。

（2）接入服务网络（ASN）

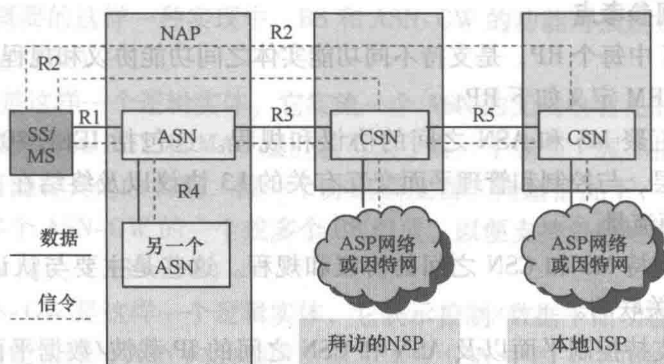


图 11.4 网络参考模型

为向 MS 提供无线电接入，ASN 支持 WiMAX 网络所需功能的一个全集。事实上，这是 MS 访问 WiMAX 网络的入口点。一个 ASN 支持如下功能：

1）无线电资源管理和与 MS 的 L2（802.16层2）连接。

2）针对 MS，合适 NSP 的网络发现和选择。

3）IP 地址分配和与 MS 的 L3 连接。

4）为用户会话建立，支持 AAA。

5）QoS 策略管理。

6）对 ASN-ASN 和 ASN-CSN 隧道的支持。

7）为支持 MS 移动性，提供 ASN 锚点的移动性、CSN 锚点的移动性、寻呼和位置管理。

（3）连接服务网络（CSN）

一个 CSN 提供所有功能，向用户提供 IP 连接能力。它由几个网元组成，如路由器、AAA 代理/服务器、一个用户数据库、互联网关设备和 DHCP/域名服务器（DNS）。一个 CSN 支持如下功能：

1）因特网接入和 AAA 服务。

2）在用户订阅概要上，对用户实施策略和接纳控制。

3）用户会话的 IP 地址和会话参数分配。

4）对 ASN-CSN 隧道的支持。

5）CSN 间漫游隧道的支持。

6）ASN 间移动性的支持。

7）连接到 WiMAX 服务，如基于位置的服务、IP 多媒体服务、对等服务和就绪提供。

注意，采用逻辑实体识别的每项功能，可在单台物理设备或分布在多台物理设备间加以实现。只要这些实现满足 WiMAX 网络规范的功能和互操作性，就允许所有这些实现。

2. ASN 间参考点

一个 NRM 中每个 RP，是支持不同功能实体之间功能协议和规程的一个逻辑接口。WiMAX NRM 定义如下 RP：

1）R1：汇聚 MS 和 ASN 之间的协议和规程。这包括 IEEE 802.16e 规范的 PHY 和 MAC 层、与控制和管理平面交互有关的 L3 协议以及终结在 ASN 或 MS 的载波/数据平面流量。

2）R2：支持 MS 和 CSN 之间的协议和规程。这些是主要与认证、授权和 IP 主机配置管理关联的。

3）R3：支持控制平面以及 ASN 和 CSN 之间的 IP 载波/数据平面协议。这些包括 AAA 支持、策略实施、移动管理和隧道（在 ASN 和 CSN 之间以隧道方式传输数据所必要的）。

4）R4：支持 ASN 之间的控制以及载波/数据平面规程和协议，如无线电资源管理（RRM）、ASN 间的 MS 移动性和一个 MS 的空闲模式寻呼。不管各 ASN 内部如何配置，R4 也可作为任何成对 ASN 间的一个互操作 RP。

5）R5：支持控制和载波/数据平面协议，这些是支持由一个本地 NSP 操作的 CSN 和由一个拜访 NSP 操作的 CSN 之间漫游所需的协议。

注意，RP R1 ~ R4 的组合体，支持在 MS、一个或多个 ASN 和锚定 ASN 的各 CSN 中所产生功能间的互操作性。

3. ASN 逻辑实体

WiMAX NRM[2] 将一个 ASN 定义为实现 WiMAX 无线电接入服务的一个灵活的、可互操作的框架。换句话说，一个 ASN 是与网络接入服务相关联的功能实体和协议的一个逻辑聚合体。由此，将一组 ASN 功能映射到各种网元［如一台 BS 和一台 ASN 网关（ASN-GW）］的不同方式，产生了一组 ASN 概要。例如，一种这样的实现可能是分解成一个或多个 BS（附接到一个或多个 ASN-GW）的一个 ASN，如图 11.5 所示。

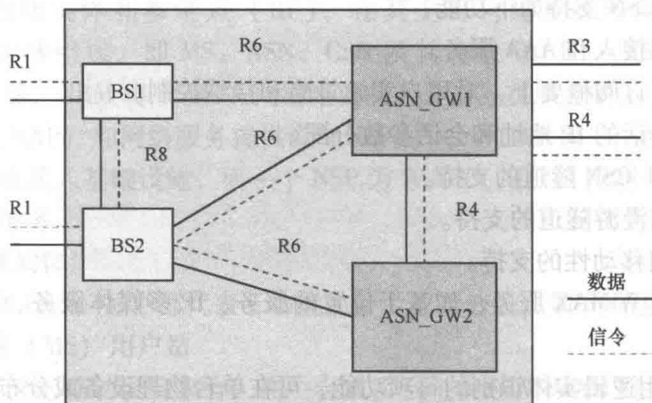


图 11.5　带有 BS 和 ASN-GW 实体的 ASN 概要

在 ASN 概要的这样一种实现中，BS 和 ASN-GW 的功能可描述如下。

(1) 基站（BS）

一台 BS 是这样一个逻辑实体，它实施一个 ASN 的无线电有关的功能。它可要求实现 IEEE 802.16e PHY 和 MAC 层。它可表示有一个或多个频率指派的一个或多个扇区。它可能有其他的实现，像一个分组调度器。典型情况下，一个 ASN 可由连接一个或多个 ASN-GW 的一个或多个 BS 组成，以便支持负载均衡和冗余。

(2) ASN 网关（ASN-GW）

一个 ASN-GW 是这样一个逻辑实体，它表示控制/数据平面功能实体的一个聚集体，这些功能实体与 ASN 内的相应功能实体或 CSN 中的一个功能实体或另一个 ASN 中的一个功能实体配成对。它也实施载波/数据平面路由和桥接功能。也存在分解 ASN-GW 功能的其他方式。

4. ASN 内参考点

下面是就 ASN 配置概要而言，在一个 ASN 内定义的各 RP：

1）R6：包括在 BS 和相关联 ASN-GW 之间的所有控制/载波平面协议。控制平面由 QoS、安全和移动管理相关的协议，如寻呼以及数据路径建立和释放，包括 RRM。载波平面表示 BS 和 ASN-GW 之间的 ASN 内数据路径。

2）R7：一个可选的 RP。如果得到支持的话，则它由 ASN-GW 内的控制平面协议组成，用于在 R6 上所涉及两个功能组之间的 AAA 和策略协调。

3）R8：为通过在各 BS（在某个 BS 的切换中涉及）之间 MAC 语境和数据的直接或快速传递而确保快速和无缝切换，在各 BS 之间的一个可选 RP。

11.4　IPv6 和 WiMAX

在过去十年间无线业界的飞速增长，支持多媒体服务和因特网接入，利用 IP 作为传输和路由分组的方法，产生了许多无线设备［个人数字助理（PDA）、蜂窝电话、epodes 等］汹涌充斥市场。IPv4 以 32 比特 IP 地址，不足以连接由无线业界引入的所有这些设备。为解决这个约束，设计 IPv6，替换 IPv4，这将地址空间从 32 比特扩展到 128 比特，由此提供一个巨大的地址空间。IPv6 也被认为是这样一个协议，被设计用来处理因特网的快速增加（是就对服务、移动性和端到端安全的紧迫需求而言的）。IPv6 也在 IPv4 之上添加如下重要改进，使之更加鲁棒、模块化和灵活：

1）无状态自动配置。

2）原生组播支持。

3）网络层安全，做法是在协议规范中集成 IP 安全（IPSec）。

4）通过移动 IPv6（MIPv6），支持原生移动性。

预计 IPv6 会逐步替换 IPv4，在过渡阶段这两种协议会共存多年时间。采用各

种过渡技术，缩短过渡阶段。

由 IPv6 引入的、优于 IPv4 的最具突破性的改进之一是无状态自动配置（RFC 2462），采用这种方法，在没有网络介入的情况下，IPv6 使主机能够建立自己的 IP 地址。接下来，这种 IPv6 无状态自动配置机制，使用 IPv6 的邻居发现（RFC 2461）。事实上，这些规程提供主机的即插即用联网方式，同时避免管理开销。所以，比较详细地理解这些功能特征是有益的。

11.4.1 邻居发现

邻居发现协议（NDP）（RFC 2461）定义了在一个 IPv6 子网或 L3 链路内节点之间的规程。一个 IPv6 子网是在因特网中具有一个唯一 IPv6 前缀的区域。在规程中涉及的一个节点可以是一台路由器或一台主机。邻居发现规程假定在子网内存在链路范围组播（或广播；在 IPv6 中，广播与组播进行了合并）支持。这意味着任何节点，无论它是一台路由器或一台主机，都应该能够向一个子网内的各节点发送组播帧。注意，在一个子网内的一个节点，使用一个 MAC 地址将一条分组转发到下一跳。下一跳是这样的邻居，它为流量应该发送到的目的地。下一跳可能是一台路由器或目的地本身。

NDP 使用路由器请求（RS）、路由器通告（RA）、邻居请求（NS）和邻居通告（NA）消息在子网中各节点间交换必要的控制/管理信息。这些 RA、RS、NS 和 NA 消息可以一个未指派的源地址发送到一个组播目的地址。为理清思路，这里所做的仅是有关 L3 链路的过渡评述。注意，L3 链路完全不同于 L2 链路（子网）。单个 L3 可有多个子网。如此，L3 是一项通信设施或媒介，通过它，子网间的各节点在链路层通信。现在继续在下面讨论 NDP 的一些规程。

1. 路由器发现

在子网上的各主机使用这个规程，定位驻留在一条附接链路上的各路由器。在进入一个子网时，一个节点产生自己的链路本地地址，并使用 RS 学习链路参数，如链路最大传输单元（MTU）、跳限制值，甚至路由器信息。所涉及的路由器在链路上周期性地发送一条 RA，由此节点定位到它的路由器，还有 IPv6 地址前缀。注意，包含一个前缀地址的一条 RA，由网络上的一台路由器周期性地作为一条组播消息发送。路由器不维护它所发送过的消息的记录。所以，如果在网络上存在两台路由器，则它们能够发送带有不同前缀地址的 RA。接收节点自动地得到两条 RA，并在相同接口上形成两个地址。由此，各节点使用前缀将驻留在链路上的目的地与仅通过一台路由器可达的那些目的地区别开来。这项设施仅存在于 IPv6 中，IPv4 中没有这项设施。

2. 地址自动配置

在 NDP 规程的帮助下，完成地址自动配置。进入一个网络时，一个节点自动地产生接口的地址或一个接口 ID。为做到这一点，节点采用两种技术中的任一种

技术，即随机产生（RFC 3041）和使用它的 MAC 地址（RFC 2373）。之后，节点利用路由器发现子例程得到 IPv6 地址前缀。由此，组合节点 ID 和 IPv6 地址前缀，则节点就得到一个全局唯一 IPv6 地址。

3. 地址解析

在网络上的各节点使用这个规程，确定在链路上的一个目的节点（即一个邻居节点，仅给定目的 IP 地址）的链路层（MAC）地址。

4. 下一跳判定

在网络上的各节点使用这个子例程，在流量到达目的地址前，确定流量的下一跳，或实施邻居不可达性检测（NUD）。为确定下一跳，这个子例程将一个 IP 目的地址映射到最佳邻居的 IP 地址，其中流量要发送到的目的地是这个邻居。下一跳可能是一台路由器或一个目的地本身。为检测邻居不可达性，要检查分组的存活时间（TTL）字段。对于用作路由器的各邻居，可尝试另外的默认路由器。在这种情形中，路由器和主机需要再次实施地址解析。

5. 重复地址检测

在一个子网中的一个节点，使用这个规程，确定它希望使用的一个地址，还没有为另一个节点使用。但是，一台路由器使用这个规程，建议主机使用另外一个第一跳路由器到达一个特定的目的地。

11.4.2 无状态自动配置

在 IPv6 中，一个 128 比特 IP 地址由两个标识符组成，即一个网络前缀（识别网络）和一个接口 ID（识别一个节点或主机的接口）。主机在其本身上配置接口 ID，而它从网络（通常是一台路由器，其中使用 NDP 规程）得到网络前缀。这两个标识符的组合，形成主机的一个唯一全局 IP 地址。换句话说，当一台主机第一次启动时，它在链路上发送一条请求，从一台 IPv6 路由器请求它的网络前缀。通过使用这个网络前缀的 MAC 地址或一个私有随机数使用 NDP 规程，它可自动配置一个有效的、唯一的全局 IP 地址。注意，这些 NDP 规程[10]非常依赖于低层能力，来处理组播通信。

自动配置使用的规程：

1）在进入网络时，新节点（如 A）产生一个链路本地地址，并将之分配给一个接口。链路本地地址有形式 fe80::/64。

2）节点 A 应用一个重复地址规程（NDP 的组成部分），确认所产生的链路本地地址没有为相同网络上的任何其他接口所用。

3）通过使用这个链路本地地址，节点 A 发送一条 RS（在图中图示为 NDP 规程）来请求信息。

4）一般而言，接收到 RS 消息的节点会发回一条 RA 消息，该消息有 IPv6 地址前缀。注意，所涉及的路由器也在网络上周期性地发送 RA 消息。所以，单单为

得到一个前缀地址，节点 A 发送一条 RS 是不必要的。

5）最后，通过组合前缀和链路本地地址或使用它自己的 MAC 地址，节点 A 形成一个全局 IP 地址。（欲了解详细规程，请参见有关寻址的那一章）

11.4.3 WiMAX 和自动配置

如在前面指出的，移动 WiMAX 网络规范的目标，是为大范围 IP 服务优化过的一个端到端全 IP 架构。所以，从 IPv4 过渡到 IPv6，是 WiMAX 技术的自然选择。采用这种强大的 IPv6 自动配置技术，WiMAX 可以一种即插即用方式提供大范围的服务。注意，事实上，IEEE 802.16e[15] 是基于一种点到多点架构的，其中在两台 MS 的 MAC 层没有授权直接的通信。在 MAC 层的所有通信启动和终止都在 BS 处。如此，在 WiMAX 中 MAC 层不支持组播通信。所以，考虑到从 IPv4 过渡到 IPv6，移动 WiMAX 不能支持这项强大的自动配置技术，原因是自动配置主要使用的 NDP 规程[10] 依赖于低层处理组播通信的能力。

由此，在下面各节回顾由移动 WiMAX 架构引入的主要问题，这个问题阻止部署 NDP 规程，由此阻止 IPv6 部署在这样的网络上。事实上，WiMAX 论坛迎接这项挑战，并为在 WiMAX 上部署 IPv6 提出一个模型，对这个专题有所贡献。将在下面各节回顾所有这些方面。

11.5 在 WiMAX 上部署 IPv6 的挑战

本节回顾在将 IPv6 部署到移动 WiMAX 过程中，由移动 WiMAX 架构[3] 所带来的一些挑战。如在前一节中指出的，这些挑战特别与移动 WiMAX 标准不能支持 NDP 规程有关，其中 NDP 规程与 IP 地址自动配置有关。给出这些挑战及其影响，以及可能的解决方案路线，这些见 WiMAX 论坛或 16ng 草案中的讨论。

11.5.1 组播支持

在上一节指出，由 IPv6 地址自动配置使用的 NDP 规程是基于低层组播通信的。这些 NDP 规程使用组播地址到达一个地址组。

WiMAX 架构（IEEE 802.16e）遵循采用单向连接的一种点到多点架构。如此，在下行链路帧（从一个 BS 到一个 MS）中，组播通信是可能的。在与一个 CID 相关联的上行链路帧中，一个 MS 不能使用组播寻址。但是，一个 BS 可发送与一个组播 CID 相关联的一条组播分组。与以太网不同，IEEE 802.16e 没有设施可将 IP 组播地址直接映射到层 2 组播地址。

所以，对将一个组播 IPv6 地址与 802.16 MAC 层的一个组播 CID 关联的一个规程是存在需求的[20]。结果，这要求 BS 处理组播连接。

11.5.2 子网或链路模型

一个 IPv6 子网或链路模型是在因特网中有一个唯一 IPv6 前缀的区域。在一个 IPv6 子网中，MS、BS 和接入路由器（AR）/ASN-GW 的三重组织对 NDP 规程具有重大影响，如图 11.6 所示。事实上，NDP 的一些功能变得过时了。为理解所涉及的挑战，进行了比较深入的挖掘，并据此区分两个 IPv6 前缀指派规程：每 MS 前缀和共享前缀。

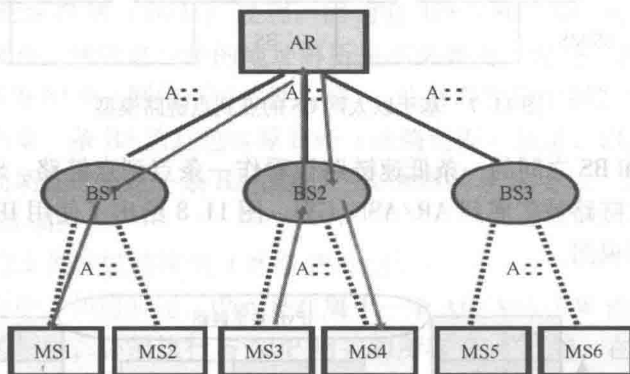


图 11.6 NWG ASN 拓扑

1. 每站 IPv6 前缀

这个 IPv6 子网或链路模型通常称作一个点到点链路模型。在这种情形中，在一个 BS 下的每台 MS 都驻留在一个不同的 IPv6 子网中。换句话说，一台 MS 和一台 ASN-GW 存在于一个 IPv6 子网下，具有链路本地范围的一个目的地址的各 IPv6 分组，仅被交付在一台 MS 和一台 ASN-GW/AR 之间的点到点链路内。

在这样一个链路模型上 IPv6 部署的一种解决方案是使用点到点协议（PPP）。因为 IEEE 802.16e 标准没有定义任何 PPP CS，所以 PPP 不能直接用在一个 IEEE 802.16e 网络上。在 IPv6 CS 子层的情形中，在这样一个链路层模型上，为在一台 MS 和一台 AR/ASN-GW 之间提供一条点到点链路，要求某种另外的机制。

另一种替代方法将是在以太网上利用 PPP[18]，做法是使用以太网 CS，以太网 CS 使用以太网上点到点协议（PPPoE）栈。图 11.7 给出了使用以太网 CS 的一个点到点架构网络实例。

2. 共享 IPv6 前缀

在这个链路模型中，一个 ASN-GW/AR 下的所有 MS（可被连接到不同 BS）都驻留在相同 IPv6 子网之中。对于这样的链路模型，有两种 IPv6 部署方案：第一种方案使用一个 IPv6 CS，而第二种解决方案使用一个以太网 CS。

（1）基于 IPv6 CS 的解决方案

在这种情形中，IPv6 层处 MS 和 AR/ASN-GW 之间的链路被看作一条共享链

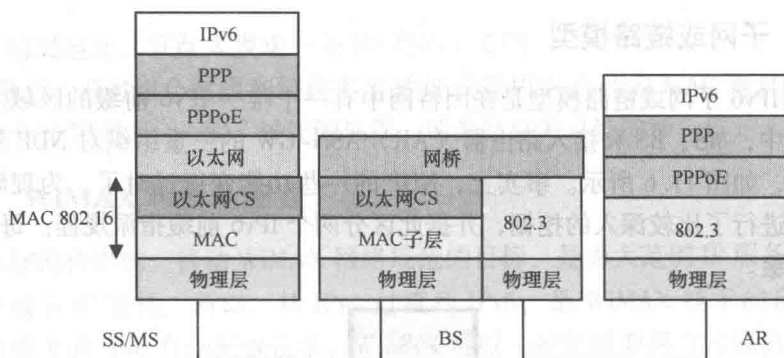


图 11.7 基于以太网 CS 的点到点链路模型

路，而 SS/MS 和 BS 之间的一条低速链路被看作一条点到点链路。SS/MS 和 BS 之间的这条点到点链路被扩展到 AR/ASN-GW。图 11.8 给出了使用 IPv6 的这个链路模型的一个高层视图。

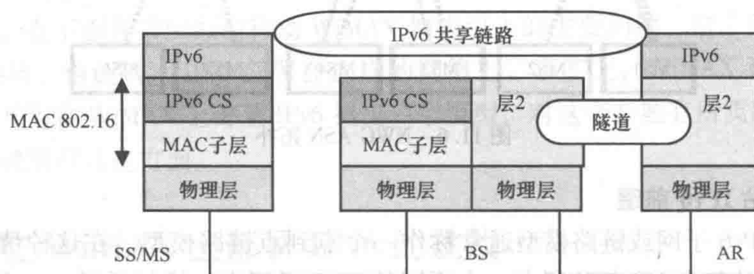


图 11.8 基于 IPv6 CS 的共享 IPv6 前缀

(2) 基于以太网 CS 的解决方案

这个模型被称为类似以太网的链路模型。它假定低层链路层提供像以太网的功能，即广播和组播。在这个模型中，假定 BS 实现桥接功能。但是，应该注意到，在以太网和 IEEE 802.16e MAC 层之间存在一点差异，如图 11.9 所示。

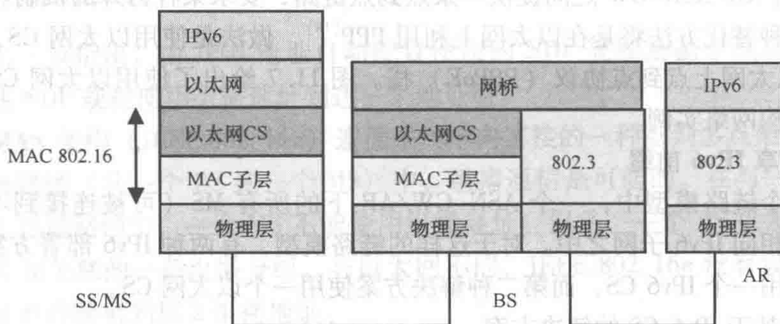


图 11.9 基于以太网 CS 的共享 IPv6 前缀

3. IPv6 功能和 CS

依据在前面各节中的讨论，注意到，取决于在移动 WiMAX MAC 层部署哪种 CS，需要利用 IPv6 功能，特别是与 NDP 规程相关的功能。注意到，依据 IPv6 子网[4]模型，存在两种类型的 CS 部署：在每站 IPv6 前缀情形中的 IPv6 CS（点到点链路模型）、在共享 IPv6 前缀情形中的以太网 CS（类似以太网的链路模型）。

（1）点到点链路模型（每站 IPv6 前缀）

在这种模型中，在前一节中看到，独立的 IPv6 前缀可被指派给每个 MS。所以不再需要重复地址检测（DAD）规程。因为在 IEEE 802.16e 帧中不使用 IEEE 802.16e MAC 缓存，所以进一步的地址解析是不必要的。但是，在存在 AR/ASN-GW 情况下，利用 NUD（网络不可达性检测）。非常需要路由器发现规程，但不清楚的是，是否需要一条 RS 消息包含源 IPv6（链路范围）地址，以便 RA 可被发送到源。同样为周期性地发送一条 RA 消息，AR/ASN-GW 需要以一种单播方式显式地发送到每个 SS/MS。

（2）类似以太网的链路模型（共享 IPv6 前缀）

在这种情形中，如前所述，IPv6 是在属于一个 AR/ASN-GW 的所有 MS 之间共享的。在这种情形中，必须执行与 NDP 相关的所有 IPv6 功能。在第一种情况中，因为在以太网地址和 IPv6 地址之间存在映射，所以要求地址解析过程。如此，当 MS 不在其网络中时，BS 使用这个封装的 MAC 地址到达目的地。同样，在这种情形中，以太网地址是以太网首部的组成部分。为得到它，非常需要 NDP 规程。针对所有 MS/SS（包括 AR/ASN-GW），为检查所有地址具有相同 IPv6 前缀，需要 NUD 规程。作为自动配置规程的组成部分，在这种情形中，DAD 是至关重要的。事实上，因为 IPv6 前缀是在 SS/MS 间共享的，所以非常有必要检查是否存在一个重复地址。明显地，路由发现规程是被严格要求，但其激活可能导致一些问题，例如能量耗散，这是移动环境中的一个重要问题。表 11.1 总结了依据两个链路模型的 IPv6 NDP 相关的功能。

表 11.1 基于链路模型的 IPv6 功能

链路模型	CS	地址解析	路由器发现	NUD	地址自动配置
点到点	IP	否	不清楚	仅检查 AR	DAD 是不简单的
以太网	以太网	是；需要地址解析	周期性地发送一条 AR	对所有 SS 和 AR 都要求	要求

4. 多链路问题

开发多数 IPv6 应用或协议时，使用 TTL 和跳数确定一条分组的范围。无论何时一条分组到达一台路由器时，在分组被转发到下一跳之前，都要将这两个参数字段做减 1 处理，由此改变了分组的范围。事实上，对于 TTL = 1，分组仅有本地范围。IPv6 协议的这个性质可导致多链路子网问题[8]，而在 WiMAX 上部署 IPv6，要

考虑IPv6前缀链路模型。这种现象可解释如下。当各MS位于不同链路上时，假定一个BS为每个MS指派独立的连接。在这种情况中，为处于不同链路上的各SS/MS分配一个共享的IPv6前缀，这意味着目的地为这些MS的所有分组，都必须经过AR/ASN-GW，原因是在两个MS之间不可能有直接通信。当AR/ASN-GW将TTL减1时，这将导致一个多链路子网问题，由此改变了分组的范围。如果在BS上或BS和ASN-GW之间没有实现任意桥接功能，则对IPv6 CS层和以太网CS层都存在这个问题。

5. SS/MS到WiMAX网络的连接

在WiMAX上部署IPv6，需要考虑一个SS/MS如何连接到网络以及如何自动配置它的IPv6地址。在WiMAX中，当一个SS/MS进入网络时，它得到三个CID连接，以便建立一个全局配置。第一个CID用于传递短的、灵敏的MAC和无线电链路控制消息。第二个CID是主要的管理连接，通过这条连接，在SS/MS和BS之间交换认证和连接建立消息。第三个CID用于辅助管理连接，管理连接部署诸如DHCP的服务。由IPv6观点看，不清楚这些CID中的哪个CID处理NDP规程。因为需要交换三条不同的管理消息，像DAD、RS和RA，所以要求采用一个不同的CID。

11.6 对建议解决方案的讨论

为解决在WiMAX上部署IPv6的挑战，16ng组和WiMAX论坛的NWG识别出要解决的一些问题，并提出克服它们的一些解决方案。在本节中，就有关WiMAX上部署IPv6方面，给出当前存在的解决方案和一些开放问题。

11.6.1 组播支持

由前一节的讨论，明显的是，在WiMAX上部署IPv6时，对组播通信存在强烈需要。所以人们期望的是在移动WiMAX（IEEE 802.16e）网络中要有IPv6组播支持。在参考文献中有许多解决方案，将在下面讨论这些方案中的一些方案。

1. 支持组播CID

这个解决方案是由WiMAX论坛的16ng组提出的[20]。它使用一个专用的CID用于组播，即组播连接标识符（mCID）[20]。在这个方面，原始的CID字段由mCID格式替换。mCID格式的细节如图11.10所示。mCID由三个字段组成：mCID前缀、CS和范围。

字段长度	6字节	1字节	4字节
字段	mCID前缀	CS	范围

图11.10 组播CID

mCID 前缀用来指明一条组播分组是内嵌于 802.16e 帧内的。CS 字段确定 CS (1 指明 IPv6 CS，而 0 指明以太网 CS)，而范围指明在 IEEE 802.16e 帧中内嵌分组的范围。但就 mCID 如何初始和在 SS/MS（参与到组播中）间如何分配，草案保持沉默（不发表意见）。

2. MRP 层

这篇文章[9]提出使用 IPv6 层（或以太网）和 IEEE 802.16e MAC CS 之间的一个中间层，即组播中继部分（MRP）。这个 MRP 层专用于 NDP 规程，并需要在 SS/MS 和 AR/ASN-GW（见图 11.11）中引入。

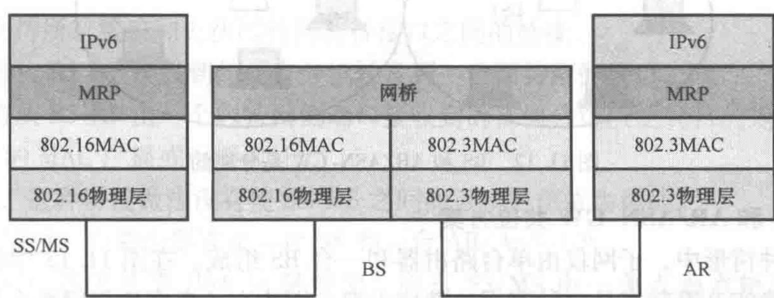


图 11.11 MRP 架构

当在 SS/MS 的 IPv6 层发出一条组播分组时，MRP 层捕获这条分组，并发送到 AR/ASN-GW。AR 的 MRP 层检查映射表，并通过重复的单播传输将这条组播分组发送到这个地址中所涉及的 SS/MS。采用这种方法的主要问题是，组播分组被 AR/ASN-GW 处的重复单播传输所替换，由此没有利用组播的优势。

3. 组播模拟

这篇文章[10]引入组播模拟的一个规程。它提出，在一台 BS 中引入组播分组处理所有相关的规程。为优化空中资源的使用，一台 BS 可施用有关执行组播规程的选择性判定。

迄今为止讨论的所有方法，在合理的条件下工作良好，但它们不能解决 MAC CS 的多链路问题和异构性，特别当一个 SS/MS 本地选择一个 CS 时更是如此。

11.6.2 BS 和 AR/ASN-GW 接口

至今，存在许多架构性的解决方案，但就一台 BS 和一台 AR/ASN-GW 之间的接口，却没有清晰的描述。WiMAX 论坛提出一种架构性解决方案，将在下面加以回顾。这篇文章[19]讨论了许多架构性的解决方案，但下面是两个主要的解决方案，即 BS 和 AR/ASN-GW 分离方案，以及 BS 和 AR 共处于一台设备中的方案。

1. BS 和 AR/ASN-GW 分离方案

在这种情形中，BS 可作为在链路层处 IEEE 802.16e 网络和 AR/ASN-GW 的一个网关。连接 BS 和 AR 的链路可以是基于以太网的。在图 11.12 中给出了这个架

{全文地址} http://202.206.108.43:8099/13/diskqrp/qrp46/04/

【作　者】（美）阿索克 K.塔卢克达尔（Asoke K.Talukder），（美）努诺 M.加西亚（Nuno M.Garcia），（美）加雅希尔莎 G.M.（Jayateertha G.M.）主编
【形态项】387
【出版项】北京：机械工业出版社，2016.03
【ISBN号】978-7-111-52645-2
【原书定价】99.00
【中图法分类号】TN915.04

封面
书名
版权
前言
目录
第1章　全IP网络：导论
1.1　各代因特网
1.2　无线因特网
1.3　全IP网络
参考文献
第2章　IPv6中的寻址和路由
2.1　引言
2.2　寻址
2.2.1　寻址概述
2.2.2　单播寻址
2.2.3　组播寻址
2.2.4　任意播地址
2.2.5　主机和路由器的地址
2.2.6　地址块分配
2.2.7　单播或任意播地址指派规程
2.3　IPv4到IPv6转换
2.3.1　转换技术
2.3.2　双栈方法
2.3.3　打隧道（协议封装）方法
2.3.4　转换方法
2.4　路由
2.4.1　网络架构
2.4.2　路由核心知识
2.4.3　路由协议
2.5　多穴连接
2.5.1　因特网结构
2.5.2　主机多穴连接
2.5.3　站点多穴连接
2.6　移动性
2.6.1　移动IPv4
2.6.2　移动IPv6
参考文献

第3章　因特网云内部的路由
3.1　网络、因特网和层
3.1.1　层相互作用
3.1.2　因特网基础设施（因特网云内部是什么）
3.2　网络和路由
3.2.1　IP寻址
3.2.2　网络和流量：电路和分组（数据报）交换
3.2.3　网络设备
3.2.4　网络流量路由
3.3　路由协议
3.4　主要路由协议
3.4.1　路由信息协议
3.4.2　内部网关路由协议
3.4.3　增强的内部网关路由协议
3.4.4　开放最短路径优先
3.4.5　路由器类型
3.4.6　边界网关协议
3.5　组播路由
3.5.1　组播编址指派
3.5.2　组播组
3.5.3　组播树
3.5.4　组播转发
3.5.5　组播路由算法
3.5.6　组播组成员关系协议
3.5.7　组播路由协议
3.6　虚拟路由器和负载均衡
3.7　基于策略的路由
3.7.1　引言
3.7.2　策略路由
3.7.3　策略路由结构
3.7.4　实现策略路由
3.8　路由器和交换机：平台架构
3.9　安全管理
3.9.1　OSPF
3.9.2　BGP
3.10　电信和公众网络：交换和路由
3.11　无线、移动、自组织和传感器网络中的路由

3.12　网络、复杂性的本质和其他创新
参考文献
第4章　全IP网络：移动性和安全性
4.1　引言
4.2　移动IP
4.2.1　发现
4.2.2　注册
4.2.3　打隧道
4.3　IPv6的移动IP
4.3.1　移动IPv6的基本操作
4.3.2　移动IPv4和移动IPv6之间的差异
4.3.3　移动IPv6安全
4.3.4　移动IPv6中的切换
4.3.5　3G CDMA网络之上移动IPv6中的切换
4.4　IP网络中的安全
4.4.1　IPsec如何工作
4.4.2　IPsec中的各元素
4.4.3　外发IP流量处理（保护到未保护）
4.4.4　进入IP流量处理（未保护到保护）
4.5　融合网络中的认证、授权和计费
4.5.1　Diameter
4.5.2　移动IPv6中的AAA
4.5.3　一个融合的移动环境的安全框架
4.5.4　3GPP安全
参考文献
第5章　转换扩展的家庭：步向基于IP的异构以用户为中心融合环境的下一步骤
5.1　引言
5.2　新的全IP家庭场景
5.2.1　高清多媒体服务蓬勃发展
5.2.2　通信流的重新分发
5.2.3　IP家庭中的服务重新分发
5.2.4　全IP家庭骨干的容量
5.3　家庭（全IP）骨干
5.3.1　IP作为家庭骨干网络的关键实体
5.3.2　家庭网络相关的联网技术
5.3.3　联网技术总结

5.4　家庭网关
5.5　桥接各项技术：步向全IP基础设施
5.5.1　桥接全IP融合架构
5.5.2　不需新导线作为全IP基础设施的一种解决方案
5.5.3　物理媒介和协议融合
5.6　全IP家庭网络基础设施之上的服务
5.6.1　扩展家庭之上的随身使用四重播放服务
5.6.2　e-健康应用
5.6.3　隐私、安全和用户概要
5.7　扩展家庭网络
5.7.1　全IP融合网络上的垂直和水平传输
5.7.2　全IP扩展家庭基础设施中的QoS
参考文献
第6章　无线车载网：架构、协议和标准
6.1　引言
6.2　实施主动安全
6.3　车辆网络架构
6.3.1　智能车辆
6.3.2　路侧单元和车载单元
6.3.3　车辆通信
6.4　车辆应用
6.4.1　安全相关的应用
6.4.2　非安全（便利性、舒适度）应用
6.5　车载标准
6.5.1　陆地移动的通信接入
6.5.2　汽车到汽车（汽车间）通信联盟
6.5.3　车辆环境中的无线接入
6.6　无线车载网络中的挑战
6.7　小结
参考文献
第7章　下一代IPv6网络安全：步向自治的和智能的网络
7.1　引言
7.1.1　背景
7.1.2　下一代IPv6网络
7.1.3　本章结构
7.2　相关工作、工具和协议
7.2.1　入侵检测／防御系统概述

7.2.2　监测网络流量
7.2.3　分组采样和流采样
7.2.4　深度分组检测
7.3　IPv6网络安全和用户剖析的智能
7.3.1　分析器
7.3.2　中心式服务器
7.4　小结
参考文献
第8章　物联网
8.1　物联网：新型因特网
8.1.1　引言
8.1.2　社会影响
8.2　物联网的特点
8.2.1　典型的LoWPAN节点的特点
8.2.2　LoWPAN
8.3　实现物联网的标准
8.4　用于物联网的协议层
8.5　用于物联网的IEEE 802.15.4　——PHY和MAC
8.5.1　868/915MHz频带
8.5.2　2.4　5GHz ISM频带
8.5.3　网络拓扑
8.6　IPv6
8.7　6LoWPAN：在无线个域网之上传输IPv6
8.7.1　LoWPAN帧格式和交付
8.7.2　一个6LoWPAN的邻居发现
8.7.3　6LoWPAN中的IPv6地址自动配置
8.7.4　首部压缩
8.7.5　6LoWPAN网状路由
8.7.6　LoWPAN广播
8.8　传输层
8.9　应用层协议
8.10　物联网的网络架构
8.10.1　自治LoWPAN
8.10.2　具有扩展因特网连接能力的LoWPAN
8.10.3　真正的物联网
8.11　安全考虑
8.12　物联网的应用

8.12.1　智能电网
8.12.2　工业监测
8.12.3　结构监测
8.12.4　保健
8.12.5　连接的家庭
8.12.6　远程测量
8.12.7　农业监测
参考文献
第9章　6LoWPAN：采用IPv6互联物体
9.1　引言
9.2　传感器节点
9.3　IEEE 802.15.4标准
9.4　6LoWPAN
9.4.1　6LoWPAN适配层
9.4.2　6LoWPAN路由
9.4.3　网状网之下路由
9.4.4　路由之上路由
9.4.5　6LoWPAN地址指派
9.4.6　6LoWPAN首部压缩
9.4.7　6LoWPAN分片
9.4.8　6LoWPAN邻居发现
9.5　6LoWPAN实现
9.5.1　TinyOS
9.5.2　ContikiOS
9.6　小结
致谢
参考文献
第10章　光纤上的IP
10.1　引言
10.2　封装中的网络数据
10.3　为什么需要帧
10.4　IP和光网络
10.5　WDM网络中的控制
10.6　IP域中的分组汇聚
10.7　全IP光突发交换网络
10.8　小结
参考文献

第11章　WiMAX上的IPv6
11.1　引言
11.2　WiMAX技术概述
11.2.1　物理层
11.2.2　MAC层
11.3　WiMAX网络架构
11.4　IPv6和WiMAX
11.4.1　邻居发现
11.4.2　无状态自动配置
11.4.3　WiMAX和自动配置
11.5　在WiMAX上部署IPv6的挑战
11.5.1　组播支持
11.5.2　子网或链路模型
11.6　对建议解决方案的讨论
11.6.1　组播支持
11.6.2　BS和AR/ASN-GW接口
11.6.3　AR/ASN-GW和NDP规程
11.6.4　子网模型
11.6.5　移动性
参考文献
封底